FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

TweeProfiles3: visualização de padrões espacio-temporais no Twitter

André Filipe do Couto Maia

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

Orientador: Prof. Carlos Soares

Co-Orientador: Prof. Pedro Abreu

January 26, 2015

U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

MIEEC - MESTRADO INTEGRADO EM ENGENHARIA ELETROTÉCNICA E DE COMPUTADORES 2014/2015

A Dissertação intitulada

"TweeProfiles3: Visualização de Padrões Espácio-Temporais no Twitter"

foi aprovada em provas realizadas em 12-02-2015

o júri

Presidente Professor Doutor Jaime dos Santos Cardoso
Professor Auxiliar do Departamento de Engenharia Eletrotécnica e de Computadores da Faculdade de Engenharia da Universidade do Porto

Professora Doutora Paula Viana
Professor Adjunto Departamento de Engenharia Eletrotécnica do Instituto Superior de Engenharia do Porto

Professor Doutor Carlos Manuel Milheiro de Oliveira Pinto Soares
Professor Associado do Departamento de Engenharia Informática da Faculdade de Engenharia da Universidade do Porto

O autor declara que a presente dissertação (ou relatório de projeto) é da sua exclusiva autoria e foi escrita sem qualquer apoio externo não explicitamente autorizado. Os resultados, ideias, parágrafos, ou outros extratos tomados de ou inspirados em trabalhos de outros autores, e demais referências bibliográficas usadas, são corretamente citados.

Autor - André Filipe do Couto Maia

Faculdade de Engenharia da Universidade do Porto

Resumo

Com o advento das redes sociais, uma grande quantidade de dados do utilizador foi gerada. Desde há alguns anos, os investigadores e as empresas aperceberam-se do valor associado a estas enormes quantidades de dados e dando origem ao desenvolvimento de algoritmos e ferramentas para extrair padrões e para os usar.
O TweeProfiles é uma ferramenta de clustering que permite analisar os dados dos tweets sob múltiplas dimensões: espacial, temporal, conteúdo e social. O objetivo foi o desenvolvimento de uma aplicação web para identificação de padrões em mensagens no Twitter. Até agora, foram desenvolvidas duas extensões para este projeto, TweeProfiles2 e Olhó-Passarinho. Ambos realizam agrupamento sobre os dados do Twitter, mas com características únicas: o TweeProfiles2, processa dados em tempo real, e Olhó-Passarinho que acrescentou à dimensão de conteúdo as imagens.
Todo o trabalho realizado no TweeProfiles até esta dissertação foi essencialmente técnico e científico com o objetivo de conceber e prototipar soluções para os desafios envolvidos. Por esta razão, ainda não foi possível fazer uma avaliação da ferramenta numa aplicação do mundo real. O objectivo desta dissertação é dar um primeiro passo neste sentido, em que o domínio de aplicação é o jornalismo.
Para atingir este objectivo, começámos por fazer um levantamento dos requisitos específicos desse domínio de aplicação. Com base nesses requisitos e numa avaliação do estado do TweeProfiles2, foram realizadas algumas tarefas de manutenção e desenvolvimento para consolidar a ferramenta. Para além da resolução de alguns problemas na implementação existente, esta consolidação teve por objetivos melhorar tanto a recolha de dados como a interação com o utilizador, aspetos essenciais para podermos passar o sistema para produção. O mecanismo de recolha de dados foi substituido pelo SocialBus. O SocialBus é uma ferramenta de recolha, processamento e armazenamento de dados de redes sociais, em particular do Twitter. Em relação à interação com o utilizador, foi realizado um inquérito com pessoas experientes em jornalismo, a fim de entender as necessidades e desejos de uma plataforma como o TweeProfiles3. Foi implementado um processo de visualização adequado para o fluxo de dados, usando vários widgets para melhor representar toda a informaçao. O sistema foi desenvolvido tendo em conta o estado da arte dos projetos nesta área e os resultados obtidos do estudo do utilizador, para podermos dar um passo em frente.
A utilidade da ferramenta desenvolvida para o jornalismo foi avaliada com base num teste de usabilidade. Apesar de ter sido realizado com um pequeno conjunto de utilizadores, este teste serviu para atingir os objetivos do projeto, nomeadamente, fazer uma primeira avaliação da utilidade da ferramenta TweeProfiles numa aplicação real. Os resultados obtidos permitiram não só as suas potencialidades como questões a melhorar.

Abstract

With the advent of social networking, a lot of user-specific, voluntarily provided data has been generated. A few years ago, researchers and companies noticed the value that lied within those enormous amounts of data and developed algorithms and tools to extract patterns from those data and to use them.
TweeProfiles is a clustering tool that analyses tweets over multiple dimensions: spatial, temporal, content and social. The goal was to develop a web application to identify patterns in Twitter posts. So far, there have been two extensions to this project, TweeProfiles2 and Olhó-Passarinho. Both perform clustering over Twitter data but with unique features: TweeProfiles2 processes real-time data and Olhó-Passarinho integrated the analysis of images in the content dimension.
All work in TweeProfiles, up to this dissertation, was essentially technical and scientific in order to design and prototype solutions for the challenges identified. For this reason, it has not yet been possible to evaluate the tool in a real-world application. The aim of this work is a first step in this direction, in which the application domain is journalism.
To achieve this goal, we started by making a survey of the specific requirements of this application domain. Based on these requirements and an assessment of the state of TweeProfiles2 some maintenance and development activities were carried out to consolidate it. In addition to solving some problems, this consolidation aimed to improve both the data collection process as well as user interaction, essential aspects in order to switch the system to production. The data collection mechanism has been replaced by the SocialBus platform. SocialBus is a tool for the collection, processing and storage of data from social networks, namely Twitter. Regarding the interaction with the user, an investigation with a small group of experienced people in journalism were surveyed, in order to understand the needs and desires for a platform such as TweeProfiles3. A visualization process suitable for data streaming was designed, using multiple widgets to better represent all the information. The system was developed taking into account the state of the art projects in this area and the results obtained from the user study, in order to move the tool one step forward.
The usefulness of the developed tool for journalism was evaluated based on a usability test. Although it was carried out with a small set of users, it was sufficient to achieve the objectives of the project. In particular it enabled a first assessment of the usage of TweeProfiles in a real application, identifying not only its potential as well as issues that need improvement..

Agradecimentos

Começo por agradecer ao meu orientador, Carlos Soares, pela oportunidade de realizar este projecto em colaboração com o SAPO, pela orientação global do trabalho e propostas de soluções durante todo o desenvolvimento.
Agradeço igualmente ao meu co-orientador, Pedro Abreu, por toda a ajuda durante todo o projecto. Sem estas duas pessoas, esta dissertação não teria tido pernas para andar.
Agradeço também ao (Dr.) Tiago Cunha. Um sincero obrigado por toda a ajuda, disponibilidade e apoio fornecido durante estes últimos meses.
Uma palavra de apreço a todos os que me acompanharam durante esta (longa) jornada. A todos aqueles, que de capa e batina, faziam um ano normal, num ano melhor, para todos nós, criando memórias que nos vão acompanhar para sempre e amizades que não se vão perder.
Quero também agradecer ao SAPO Labs por toda a disponibilidade para ajudar no projecto.
Um obrigado aos meus companheiro do Pub 1808, por sempre se interessarem na dissertação e apoiarem em todos os momentos.
Por último, mas com mais importância, quero agradecer à minha famílias, especialmente aos meus pais, por todo o apoio que me deram e por acreditarem sempre no meu sucesso, mesmo quando as coisas estavam mais complicadas.

À magi, Paula Vieira, um agradecimento (ainda mais) especial, que sem ela, esta jornada nunca teria sido a mesma. Foi (e é) a minha maior companhia e o meu maior apoio, em todos os bons ou maus momentos, a minha maior força para nunca desistir.

André Maia

"Some men aren't looking for anything logical, like money. They can't be bought, bullied, reasoned or negotiated with. Some men just want to watch the world burn."

Alfred Pennyworth

Contents

List of Figures

List of Tables

Symbols and Abbreviations

3D	Three dimensional
2D	Two dimensional
DB	Database
AI	Artificial Intelligence
API	Application Programming Interface
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
JPN	Jornalismo Porto Net

Chapter 1

Introduction

Social networks have a major impact nowadays. Their influence is felt in several aspects. Online social networks present a variety of social media services which have achieved a huge importance in social life as well as in marketing strategies as they "have been regarded as a timely and cost-effective source of spatio-temporal information" [4]. One business which has been significantly affected is journalism. Social networks can be used to change the way journalists are able to take the pulse of the trending themes or topics that are currently being talked about.

Recently, it acquired a new meaning in information science which is "a dedicated website or other application which enables users to communicate with each other" [11]. The massive adhesion and the number of platforms that provide social interaction lead to a growth in the data stored within these services. This data has been used by many investigators as a source of information [12, 13, 14].

Unlike what happens with other social networks like Facebook[1] and Linkedin[2] that use a network of bi-directional communication, Twitter uses an asymmetric infrastructure where there are "friends" and "followers". Assuming that you are a user of Twitter, the "friends" correspond to the accounts of people that you follow and "followers" correspond to the accounts of people who follow you [15]. Twitter is one of the top social networks, both in popularity (worldwide public awareness) and monthly active users (around 250 million [16]). It is a starting point for our ongoing work because it is the most widely used microblogging application, with the number of 'tweets' reaching 110 million per day in January 2011 and is still escalating [7].

The initial TweeProfiles [17, 18] is focused on identifying profiles on data collected [information and extraction have specific meanings in this context which may be misleading] from Twitter. The data is processed over 4 dimensions (spatial, temporal, social and content) using Data Mining techniques. The tool enables the visualization of the results of the clustering algorithm.

Since it lacked the ability to produce real-time visualizations of the evolution of the data stream, as well as the ability find patterns in the images attached to tweets, two extensions were developed:

[1] More information at: https://facebook.com

[2] More information at: https://linkedin.com

TweeProfiles2 [10] and Olhó-Passarinho [19]. Both have the same goal as TweeProfiles, to identify profiles on multiple dimensions. Due to an unsatisfactory handling of the social dimension in the original tool, these extensions instead of processing the data over four dimensions, only use three: spatial, temporal and content. TweeProfiles2 replaced the original batch clustering algorithm with a stream clustering algorithm, enabling the use of real-time data. However, the system was not fully implemented and still used static data to perform clustering, and the results obtained were as good as they were previously in TweeProfiles. Olhó-Passarinho introduced a new set of results, since it was the first TweeProfiles version using images as part of the content of tweets.
Until now, all work done was essentially technical and scientific, in order to design and prototype solutions to the challenges involved. For this reason, it has not yet been possible to evaluate the visualization platform in a real-world application.. Visualization environments and techniques provide an important function in communicating urban research and support collaborative endeavours [20]. Big advances in digital technologies and the wide availability of the Internet enables producing, manipulating, and sharing vast digital data resources, many of which contain geospatial references [21]. With more digital data becoming increasingly available there are novel ways urban researchers can now explore urban space and place in novel ways, supported by a wide array of visualization tools and techniques. Some of these tools use Data Mining techniques to generate knowledge. Some of the main functionalities already implemented in this platform are the visualization of the location where a tweet is posted, its content and different size and location of clusters.

1.1 Motivation

Over the years, Twitter has become one the major social networks to share ideas and information. It contains countless data regarding people's interests, due to the large number of users. This makes Twitter a perfect service to collect data, providing researchers the necessary information for the development of data analysis and knowledge extraction tools.
Since the majority of information shared is in text format, Twitter posts have been the source of data used by tools such as TweeProfiles [17] [18] and TweeProfiles2 [10]. However, it allows sharing pictures directly or through services such as Twitpic[3] or Instagram[4]. These images can also be used for data analysis, due to the fact that in some cases it may be complementing the text or even replacing it. The analysis of visual information is important and enabled the development of the TweeProfiles' extension Olhó-passarinho [19].
Our motivation lies with the design of an extension for TweeProfiles2. One big contribution with this project is the development of a visualization system with an interactive representation of the profiles as well as the messages themselves, connected with other platforms from SAPO Labs [5]

[3]More information at: https://twitpic.com
[4]More information at: https://instagram.com
[5]More information at: http://labs.sapo.pt

and the evaluation of this tool in a real-world application, as well as the integration with a real-time data extraction platform named SocialBus[6] [22].

1.2 Objectives

This dissertation aims to evaluate the tool in a real-world application by developing an extension of TweeProfiles2 . The goals are: 1) to complete the work done for TweeProfiles2 regarding the social dimension; 2) to complete the integration of TweeProfiles2 with SocialBus; 3) to develop an interactive visualization tool for displaying profiles and tweets found as well as a connection to information in SAPO[7].

1.3 Structure

This dissertation is organized as follows: Chapter 2 contains the state of the art for the scientific themes related to this project, namely some stream clustering algorithms and distance measures for each dimension in spatio-temporal Data Mining, alongside a state of the art study on spatio-temporal visualization techniques to use in this tool.
In Chapter 3 we explain the whole architecture of TweeProfiles3, detailing the integration with Social Bus, the survey done for the system requirements and its implementation. In Chapter 4 we present the testing setup, detailing some parameters in the clustering algorithm. A few examples of the results that were obtained are presented and analysed in order to provide some validation and to illustrate the type of knowledge that can be extracted. It also presents the final survey done with some journalists where a usability test was performed. Finally, in Chapter 5 we discuss the results obtained and some of the decisions made throughout the development. The most important limitations are discussed and future work is reviewed.

[6]More information at: https://reaction.fe.up.pt/socialbus
[7]More information at: http://sapo.pt

Chapter 2

State of the Art

This chapter summarizes the results of the study done to acquire the required skills and knowledge to develop this work. It is focused primarily in visualization techniques but a brief analysis of data mining methods, is also presented. In Section 2.1, the technical aspects related to clustering processes and algorithms, and the similarity (distance) functions are exposed. In Section 2.2 we detail relevant techniques in visualization systems regarding social media. Finally, an overview of Twitter and TweeProfiles is done (Section 2.3).

2.1 Clustering

Data mining is the process of exploring large amounts of data with the goal of finding "interesting" patterns [23]. Data Mining is a multi-disciplinary field at the confluence of Statistics, Computer Science, Machine Learning, Artificial Intelligence (AI), Database Technology, and Pattern Recognition, as [24] claims. This process is defined by several tasks, depending on the problem being addressed, such as [25]:

- **Detection of anomalies (outliers / changes / deviations)** - Identify records of unusual data. Those may be errors in the data or interesting objects that exhibit different behaviour from the most typical one;
- **Mining of frequent patterns, associations and correlations** - Finds patterns that occur frequently in the data;
- **Classification** - Learns a model mapping the values of input (independent) variables with an nominal output (dependent variable or target attribute) and applies it to new data, being mainly used in tasks of forecasting;
- **Regression** - Similar to classification, except that the target variable is numeric;
- **Summary** - Compact representation of a data set, which may include description and visualization through a report;

- **Clustering** - Groups objects in subgroups, where examples from the same subgroup are similar to each other and different from objects in the other subgroups, without without receiving any information about the characteristics of the subgroups.

Since this project follows the work done in TweeProfiles2, one of its main tasks is to perform clustering on data collected from Twitter. Clustering is defined as "the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters" [26]. It falls into a set of techniques in the in unsupervised learning type of Data Mining methods. While in supervised learning, data is previously analysed and labelled and then used to train models able to find relationships between the attributes of these data with the target variable, in unsupervised learning data sets are not previously labelled. Thus the process of discovering patterns in the data only takes into account the data it received in that moment, trying to organize similar instance [27] groups. The level of similarity between objects is dependent on the distance function used. Different distance functions are used in different clustering methods, which means that different methods and algorithms applied to the same dataset may generate different clusters. A similarity measure is calculated through distance functions, which are described in section 2.1.3.

2.1.1 Stream Clustering

As it has been stated by Gama et al [2], the world of today is "a world in movement". The data sought to analyse and extract information from has new characteristics [10]:

1. Data is made available through *unlimited streams* that continuously flow, eventually at high speed, over time;
2. The underlying *regularities may evolve over time* rather than be stationary;
3. The data can no longer be considered *independent and identically distributed*;
4. The data is now often *spatially situated* as well as time situated.

In order to address the unpredictability and volume of the data that was the aim of TweeProfiles2, a special type of clustering algorithms was needed, which are designed to treat streams of data. These are known as stream clustering algorithms and vary from regular clustering algorithms in a way that they accomplish certain requirements. These requirements are [28]:

1. Compactness of representation;
2. Fast, incremental processing of new data points;
3. Clear and fast identification of outliers.

Stream clustering is a research area that recently emerged to discover knowledge from large amounts of continuously generated data. In this context, several algorithms have been proposed to perform unsupervised learning [1]. In [1], a survey of stream clustering algorithms was conducted, detailing 13 major stream clustering algorithms, including the ones detailed above. Several important aspects of stream clustering are referred and compared, namely the cluster (and/or micro-cluster) structures used by each algorithm, the temporal decay approaches, or the applications they were used in, among others [10]. Figure 2.1 shows the list of the algorithms analysed [1].

Algorithm	Data Structure	Window Models	Outlier Detection	Number of Parameters
(1) *BIRCH*	feature vector	landmark	density-based	5
(2) *CluStream*	feature vector	landmark	statistical-based	9
(3) *ClusTree*	feature vector	damped	—	3
(4) *D-Stream*	grid	damped	density-based	5
(5) *DenStream*	feature vector	damped	density-based	4
(6) *DGClus*	grid	landmark	—	5
(7) *ODAC*	correlation matrix	landmark	—	3
(8) *Scalable k-means*	feature vector	landmark	—	5
(9) *Single-pass k-means*	feature vector	landmark	—	2
(10) *Stream*	prototype array	landmark	—	3
(11) *Stream LSearch*	prototype array	landmark	—	2
(12) *StreamKM++*	coreset tree	landmark	—	3
(13) *SWClustering*	feature vector	landmark	—	5

Algorithm	Cluster Algorithm	Cluster Shape	Cluster Problem
(1) *BIRCH*	*k*-means	hyper-sphere	object
(2) *CluStream*	*k*-means	hyper-sphere	object
(3) *ClusTree*	*k*-means / *DBSCAN*	arbitrary	object
(4) *D-Stream*	*DBSCAN*	arbitrary	object
(5) *DenStream*	*DBSCAN*	arbitrary	object
(6) *DGClust*	*k*-means	hyper-sphere	attribute
(7) *ODAC*	hierarchical clustering	hyper-ellipsis	attribute
(8) *Scalable k-means*	*k*-means	hyper-sphere	object
(9) *Single-pass k-means*	*k*-means	hyper-sphere	object
(10) *Stream*	*k*-median	hyper-sphere	object
(11) *Stream LSearch*	*k*-median	hyper-sphere	object
(12) *StreamKM++*	*k*-means	hyper-sphere	object
(13) *SWClustering*	*k*-means	hyper-sphere	object

Figure 2.1: Characterization of existing stream clustering algorithms (reproduced from [1])

.

Being the points mentioned above very pertinent and posing important questions regarding the main issues with stream clustering, Pereira states in [10] that:

- If the representation of the processed data is very thorough and precise, it will also require a great deal of resources (disk space and processing power), which, with the arrival of new data, may rapidly become impractical;
- If the processing of the new data points takes a long time, it will defeat the purpose of real-time analysis, as it will reach a point where new data will have arrived before the previous batch is processed. So, usually, stream clustering algorithms are required to process new batches of data in one pass;

- As data streams are usually very volatile in nature, we must have an effective way to distinguish between what is an outlier in the current clustering model and what is a new trend that is emerging in the stream, with the latter requiring a change in the clustering model that is currently being considered.

Figure 2.2 illustrates these requirements by comparing traditional and stream data processing characteristics.

	Traditional	**Stream**
Number of passes	Multiple	Single
Processing Time	Unlimited	Restricted
Memory Usage	Unlimited	Restricted
Type of Result	Accurate	Approximate
Distributed?	No	Yes

Figure 2.2: Differences between traditional and stream data processing [2]

Since the main differences between stream and traditional clustering methods have been mentioned, a brief introduction on the clustering algorithm used in TweeProfiles2 is done.

2.1.1.1 DenStream

Described in [29], DenStream is a density-based stream clustering algorithm. This means that, in contrary to partitioning methods, which are almost always based on distances and therefore can only find spherical clusters, in density-based methods the "general idea is to continue growing a given cluster as long as the density (number of objects or data points) in the neighbourhood exceeds some threshold" [30], allowing for the formation of arbitrarily shaped clusters. Also, density-based methods do not require the *apriori* definition of the number of clusters to be found. Density areas in regular clustering algorithms are represented by all points belonging to the cluster. However this would be impractical to do when dealing with a stream, DenStream applies the concept of core-micro-clusters.
A core-micro-cluster, at time t is defined as *CMC(w,c,r)* for a group of points $X_{i_1}...X_{i_n}$ with time stamps $T_{i_1}...T_{i_n}$:

- $w = \sum_{j=1}^{n} f(t - T_{i_j}), w > \mu$ is the weight;
- $c = \frac{\sum_{j=1}^{n} f(t-T_{i_j}) \times X_{i_j}}{w}$ is the center;
- $r = \frac{\sum_{j=1}^{n} f(t-T_{i_j}) \times dist(X_{i_j},c)}{w}, r \leq \varepsilon$ is the radius, where $dist(X_{i_j}, c)$ denotes the Euclidean distance between the point X_i and the center c [29].

The temporal decay of a data point is given by the function

$$f(t) = 2^{-\lambda \times t}, \lambda > 0$$

which means that the older the point is, the less weight it has.
However, when a new data point arrives, it is unlikely that the newly created micro-clusters obey the weight constraint. So, the concepts of potential core-micro-cluster (p-micro-cluster) and outlier-micro-cluster (o-micro-cluster) are also introduced. These are defined as regular core-micro-clusters, with the only difference being the relaxation of the weight constraint, which is $w \geq \beta \times \mu$ for p-micro-clusters and $w < \beta \times \mu$ for o-micro-clusters, with $0 < \beta < 1$.
As a new data point arrives, it is merged with one of the existing p-micro-clusters. If it is not possible, because it violates the radius constraint, the next step is to try to merge it with one of the existing o-micro-clusters. If it is successfully merged, the weight of that cluster is recalculated and depending on the satisfaction of the expression $w \geq \beta \times \mu$, the o-micro-cluster is converted into a new p-micro-cluster. Otherwise, if the new data point cannot be merged with any of the existing micro-clusters, a new o-micro-cluster is created containing solely the new data point.
The weights of the older micro-clusters decay with time. Every T_p time period, those clusters whose weight is less than the threshold ξ are deleted. This threshold is given by:

$$\xi(t_c, t_o) = \frac{2^{-\lambda \times (t_c - t_o + T_p)} - 1}{2^{-\lambda \times T_p} - 1}$$

where t_c is the current time and t_o is the time of creation of the micro-cluster.

2.1.2 Clustering Twitter Data

Until now, all sections focused on techniques for data stream clustering and analysis of existing algorithms. As stated in [1] all these are in vain unless data stream clustering is useful in practice. In this section, the applicability of stream clustering is addressed, with a brief discussion of some of the existing platforms.
Table 2.1 presents several systems that apply clustering to Twitter data, in order to understand how current systems take advantage of the information extracted from this social medium. They are analysed in terms of three characteristics: the data dimensions handled, how streaming is handled and the clustering algorithm used.

	CompactMap [3]	**EventRadar [31]**	**Visual text mining using association rules [32]**	**A Latent Variable Model for Geographic Lexical Variation[33]**
Dimensions	Content	Spatial, Temporal, Content	Content	Spatial, Content
Stream handling	Online	Online	Offline	Online
Clustering algorithm	LDA (Latent Dirichlet Allocation)	DBSCAN	Extension of Apriori	K-means

Table 2.1: Analysis of different systems with Twitter data clustering

CompactMap

CompactMap [3] packs tweet clusters effectively while generating stable layouts. It achieves coherent layout by dynamically matching clusters across time, and removes overlaps using constrained multidimensional scaling [3].

EventRadar

EventRadar [31] detects events everywhere without keeping a list of locations by finding clusters of Tweets that contain the same subset of words [31]. If a cluster is found, it is considered a potential event.

Visual text mining using association rules

In [32] a framework of tools for VTM (visual text mining) that integrates multi-dimensional exploration techniques with mining are used to extract meaning from a text data set.

A Latent Variable Model for Geographic Lexical Variation

[33] presents a method for identifying geographically-aligned lexical variation directly from raw text. It takes the form of a probabilistic model capable of identifying both geographically-salient terms and coherent linguistic communities [33].

2.1.3 Distance Measure

As mentioned in Section 2.1, clustering algorithms need distance functions in order to calculate dissimilarities between objects and to group these objects by similarity. As stated in [26] "the objective function aims for high intra-cluster similarity and low inter-cluster similarity". There are multiple kinds of distance measures that can be used, depending on the data we are processing.

Numerical Distance

In this context, numerical distance functions are used to calculate both spatial and temporal distance. Spatial dimension is defined by latitude and longitude numeric values extracted from tweets. Therefore, similarity functions between numeric values must be explored.
As claimed by Han and Kamber [30], the most popular distance measure is the Euclidean Distance (also known as straight line distance). If we consider two objects, x_1 and x_2, composed of n attributes each, the Euclidean distance between them is defined as:

$$dist(X_1, X_2) = \sqrt{(X_{1_1} - X_{1_2})^2 + ... + (X_{1_n} - X_{2_n})^2}$$

Also a very well known and often used distance, the Manhattan (or block) Distance is named in such a way since it gives us a distance in blocks between two points in a city (for example, 2 blocks down and 3 blocks over is a total distance of 5 blocks). Its formula is as follows:

$$dist(X_1, X_2) = |X_{1_1} - X_{1_2}| + ... + |X_{1_n} - X_{2_n}|$$

As for the Minkowski Distance, it is no more than a generalization of both the Euclidean and Manhattan Distances:

$$dist(X_1, X_2) = \sqrt[p]{|X_{1_1} - X_{1_2}|^p + ... + |X_{1_n} - X_{2_n}|^p}, p \geq 1$$

It is also called the L_p norm because of the notation of p in the formula. When $p = 1$ it represents the Manhattan Distance (L_1 norm) and when $p = 2$ represents the Euclidean Distance (L_2 norm). Lastly, we also have the Chebysev Distance or supremum distance (also known as *Lmax* or L_∞ norm). It is a generalization of the Minkowski formula for $p = \infty$:

$$dist(X_1, X_2) = \lim_{x \to +\infty} (\sum_{j=1}^{n} |X_{1_j} - X_{2_j}|^p)^{\frac{1}{p}} = max_j |X_{1_j} - X_{2_j}|$$

This formula gives us the maximum distance between the values of each attribute of the objects. The Haversine Distance [34] is specific for the spatial dimension. It is used to calculate the great-circle distance between two points, based on their geographical coordinates (latitude and longitude). It is an approximate distance (it has a high accuracy, though) because it assumes the Earth is a perfect sphere, when in reality is slightly flattened in the poles. The Haversine formula is:

$$distSp(X_1, X_2) = 2 \times R \times \arcsin\left(\left[\sin^2\left(\frac{\theta_{X_1} - \theta_{X_2}}{2}\right) + \cos\theta_{X_1} \times \cos\theta_{X_2} \times \sin^2\left(\frac{\lambda_{X_1} - \lambda_{X_2}}{2}\right)\right]^{\frac{1}{2}}\right)$$

where R is Earth's radius, θ the point's latitude and λ the point's longitude.

In terms of temporal dimension, time is represented in $\mathbb{R}$, as opposed to the previous distances which are in R^2, thus making its difference calculation easier. For each pair of tweets t_i and t_j, the timestamp values Δ_i and Δ_j are used to compute the distance. The following equation can define the time interval:

$$dist(t_i, t_j) = |\Delta_i - \Delta_j|$$

However, any of the previous distance functions for euclidean space are applicable to the temporal dimension [17] [18].

Textual Distance

The most commonly used measure for comparing text documents is the cosine similarity, as claimed by Han [30]. However, in order to do that, those documents must be represented in the form of a numerical vector. One way of doing that is using the bag-of-words representation. It consists of associating with each term in a document, the frequency of its occurrance in that text. Alternatively, *tf-idf* (term frequency-inverse document frequency) [35] is also a commonly used document representation. Considering a tweet T_i, its text attribute as W_i and a term in that text as t_{W_i} can be defined. The term-frequency of a term can be calculated as such:

$$TF(t_{W_i}) = \frac{freq(t_{W_i})}{N}$$

where $freq(t_{W_i})$ is the absolute frequency of the term t_{W_i} and N is the size of the text.The *tf-idf* of a term can be calculated by the following formula:

$$TFIDF = TF(t_{W_i}) \times IDF(t_{W_i})$$

where $TF(t_{W_i})$ is the term-frequency of term t_{W_i} and $IDF(t_{W_i})$ is the inverse document frequency of term t_{W_i} . The inverse document frequency is a statistic that reflects the importance of a term relative to a collection of text documents. It is given by the formula:

$$IDF(t_{W_i}) = \log\left(\frac{ND}{DF(t_{W_i})}\right)$$

where ND is the number of documents in the collection and $DF(t_{W_i})$ is the number of documents that the term t_{W_i} appears in.
Let x and y be two term-frequency vectors. The cosine similarity of these vectors given by:

$$sim(x,y) = \frac{x \times y}{||x|| ||y||}$$

with $||x||$ and $||y||$ being the norm of x and y, respectively. This measure calculates the cosine of the angle between the two vectors, meaning that the closer the angle is to 90, the more unlikely they match. The value obtained will be between 0 and 1, 0 meaning that the documents have no match, and 1 meaning the documents are exactly equal. When the attributes of the vectors evaluated are binary-valued (when it is only considered whether a word appears or not in a text) the cosine similarity function can be interpreted in terms of shared attributes. A variation of the cosine similarity in this case is the Tanimoto coefficient:

$$sim(x,y) = \frac{x \times y}{x \times x + y \times y - x \times y}$$

This formula gives us the ratio of the number of shared attributes between x and y to the total number of attributes.

2.2 Visualization

An important step in the Data Mining process is the interpretation of the results. It is often based on tools to visualize both the data and the knowledge extracted. The main properties that must be verified by these tools are: the displaying of the data and temporal behaviour; showing properties of the entire displayed scene and support interaction [36]. Visualization tools in a large multidisciplinary initiative require a pragmatic yet somewhat critical review of the ways visualization can be used to represent and to analyse data [20]. In this project, the review of related work is focused on was done based on systems that collect data from Twitter. At first, we identify some of the different dimensions these systems can take place in. These dimensions, such as the geographical coordinates (spatial dimension), the timestamps (temporal dimension), the connections between users (social dimension) and the text (content dimension), provide the context between all the data in the system. We also review some of the technology environments and platforms that are available.

These present the playground and context for the interaction between the user, the data, and the visualization product. The goal is to identify the possible interaction paradigms to be supported by the visualization tools that developed in this project. Finally, we look at some techniques that can be used to visualize geospatial data. Instead of traditional visualization techniques like graphs and charts, we analyze a set of contemporary geo-visualization techniques made to improve the experience.
Besides different techniques, different types of information are displayed, enabling the design of visualization tools based on clustering and/or georeferenced data.

2.2.1 Clustering Visualization

For clustering visualization, the most common representation are graphs. The objects in each cluster are presented and the goal of assigning similar objects the shortest distance between clusters is maintained. A system developed by [32] can be seen in Figure 2.3.

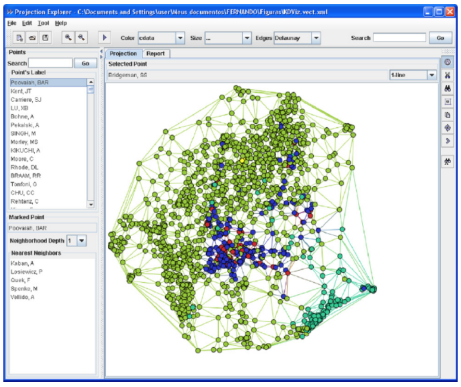
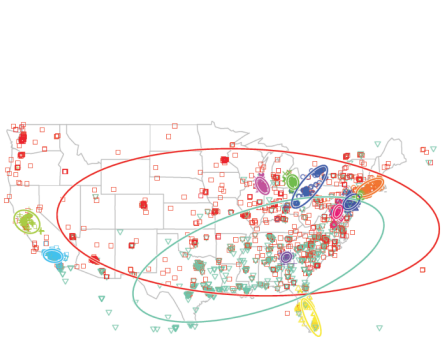
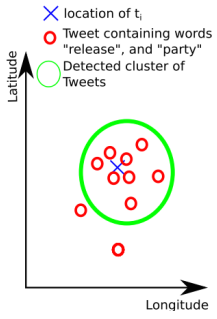


Figure 2.3: Clustering Visualization

Another clustering visualization method for a large amount of data involves assigning different colors and objects. For objects in different clusters, overlapping ellipses over the most representative objects are displayed to represent similar objects [17] [18]. This approach was applied to study geographical lexical variation (figure 2.4a [33]) and to classify events (figure 2.4b [31]).



(a) Clustering Visualization.



(b) Clustered distribution around *ti* [31].

Figure 2.4: Clustering Visualization Examples.

More recently different clustering visualization methods were designed based on a different technique. In [3] clusters are displayed in a more innovative and interactive way, enabling the user to easily identify content in each cluster. An example of this system can be seen in Figure 2.5.

Figure 2.5: Clustering Visualization of the tweets for the search term "technology" on May 16, 2013 [3].

2.2.2 Visualization Methods

Since the Twitter's API provides different types of information such as tweet ID, user ID, user name, timestamps, latitude, longitude and text, multiple visualization techniques can be used to visualize it. The following list briefly discusses some of these techniques.

- **Choropleth Maps**: A choropleth map represents aggregated measures for pre-defined geographical regions (Wright, 1938). Choropleth mapping is particularly useful in providing comparative summaries over specified geographies [20];
- **Heat Map**: Also known as density surfaces, a heat map is a visualization technique to represent the density of spatial data using Kernel Density Estimation methods [20]. It enables users to easily identify high density areas without losing the general spatial context. It is typically used to analyse geospatial data, where a sense of correlation between geographical features and other measurements is required. Heat maps are also useful in mapping temporal urban phenomena such as people and traffic flows [37];
- **Flow Map**: Flow maps display movements of objects or subjects from one place to another by means of lines or arrows [20]. The data used needs different initial and final geographical locations, such as, for example, migration patterns between regions. Although they are usually static, flow maps can be made dynamic by using a time sequence animation [20];
- **Social Map**: A social map is the cartographic, two-dimensional representation of social interactions. In systems such as [3] it is used as a representation of clusters containing similar information (in this case similar words from tweets);

- **Brushing**: Also known as multiple-linked views, it is a method for dynamic querying by direct manipulation of visual and data displays with the results being updated based on manipulation is commonly referred to as "brushing" [20]. It is usually used for exploratory data analysis [38].

Gahegan [36] claims that the main visualization techniques are: map-based, chart-based, projection, space-filling or pixel based, iconographical or compositional and hierarchical or network. This is not intended as an exhaustive review of all the relevant techniques.

2.2.2.1 Georeferenced Data Visualization

Georeferenced data involves displaying the information on a geographic representation, usually a map (such as Google Maps[1], Google Earth[2] or NASA WorldWind[3]). One other system created by Google that enables a 3D visualization of the Earth is Google Earth. With an easy interface, it allows an intuitive representation of georeferenced data..
Google Maps is a very popular 2D map visualization tool. A large number of applications were developed using their API because of its simplicity and visual appeal, including some of the tools described earlier [6, 39, 4].

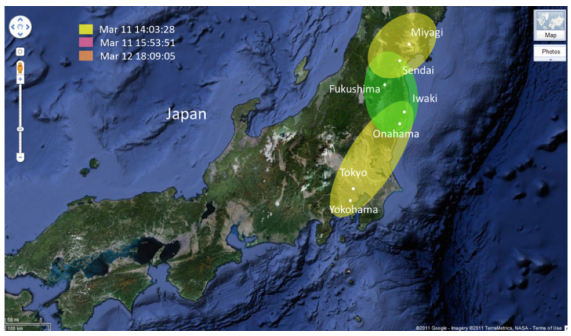


Figure 2.6: Map with event detection on Twitter [4]

Even though only visualization tools from Google have been mentioned, there are many others in this market that also provide a map API [17] [18].
Silicon Graphics International, in partnership with the University of Illinois designed a real-time visualization platform of sentiment mining on Twitter [5]. The tool adopted a heat map representation, in which each color represented a different value for the majority of positive or negative comments [17] [18].

[1] https://developers.google.com/maps/?hl=pt-pt
[2] https://developers.google.com/earth/?hl=pt-pt
[3] http://worldwind.arc.nasa.gov/

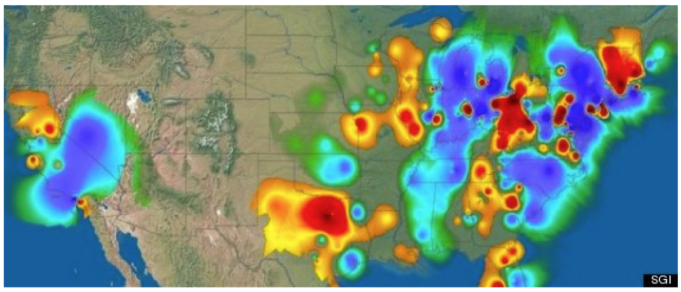


Figure 2.7: Real-time heat maps of positive and negative sentiments expressed via Twitter [5]

2.2.3 Twitter Data Visualization

An analysis of several tools for visualization of Twitter data was made. In table 2.2 some of those tools, in order to understand how current systems envision using geovisual tools can take advantage of social media data sources (namely Twitter).

	TweetPos	**SensePlace2**	**CompactMap**	**Earthquake**	**TweetDrops**	**A Software System for Data Mining with Twitter**	**EventRadar**
Dimensions	Spatial, Temporal, Content (Hashtags)	Spatial, Temporal, Content	Content	Spatial, Temporal, Content	Content	Spatial, Content	Spatial, Temporal, Content
Environment	Digital Globes	Digital Globes	N/A	Digital Globes	N/A	Digital Globes	Digital Globes
Platform	Google maps	Proprietary	N/A	N/A	N/A	Google maps	N/A
Visualization Techniques	Heat Map/ Graph & Charts	Heat Map	Social Map	Heat Map/ Graph & Charts	Rain drops	Brushing/Heat Map/ Graph & Charts	Brushing
Type of visualization	2D Area	2D Area	N/A	2D Area	2D Points	2D Area	2D Area
Type of information displayed	Density of tweets	Position and Content of tweets	Size of Clusters/ Content of tweets	Position of tweets	Content of tweets	Position and Content of tweets/ Density of tweets	Position of tweets
Stream handling	Online	Online	Online	Offline	Offline	Offline	Online
Clustering algorithm	N/A	N/A	LDA (Latent Dirichlet Allocation)	N/A	N/A	N/A	DBSCAN

Table 2.2: Differences between Visualization Systems

TweetPos

TweetPos [6] is a web service that is intended to facilitate the analytical study of geographic tendencies in Twitter data feeds. In order to improve the user's experience with the tool, TweetPos relies on visual data structures like heat maps and charts to represent the geo-spatial sources of tweets. Figure 2.8 displays an heat map example of the TweetPos tool.

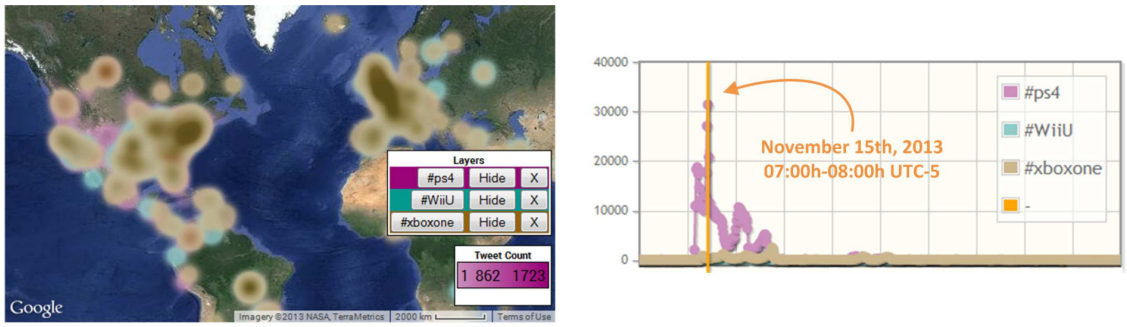


Figure 2.8: Heatmap-based as well as quantitative comparison of game console popularity from [6].

SensePlace2

SensePlace2 [7] presents a geovisual analytics approach to support situational awareness (SA) for crisis events using Twitter. It focuses on leveraging explicit and implicit geographical information for tweets and on providing visual interface methods to enable understanding of place, time, and theme components of evolving situations [7].
It is a user-centered approach, using scenario-based designs that include formal scenarios to guide and validate implementation as well as a systematic claims analysis to justify design choices and provide a framework for future testing.
It is composed by a structured survey of practitioners and the end product of Phase-I development is demonstrated through implementation of a map-based, web application initially focused on tweets but extensible to other media [7].

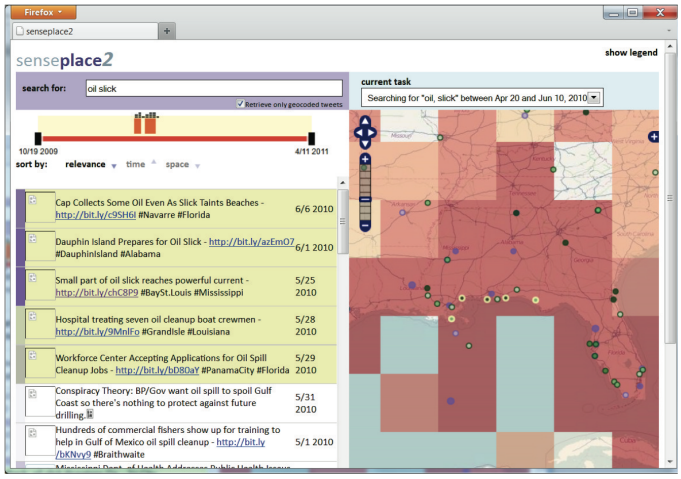


Figure 2.9: Visualization of query restricted to tweets with geo-location from [7]

TweetDrops

TweetDrops [8] is a computer-based visualization tool designed for people who have not paid

attention to sustainability in their life before. It opens up an opportunity for them to learn about energy conservation. It has two main visual components, as shown in Figure 2.10. One is the background rain drops, which represent the accumulation of energy related tweets collected from Twitter; the other is clickable foreground tweets with detailed content.

Figure 2.10: Foreground clickable tweets are displayed as large blue sentences in front of the rain drops [8].

CompactMap

CompactMap [3] is an online visual interface that packs text clusters efficiently. It achieves spatio-temporally coherent layouts by dynamically matching clusters across time, and removing cluster overlaps according to spatial proximity and constraints. CompactMap enables:

- A dynamic visualization technique that displays clusters in text streams as stable, space-efficient layouts;
- A real-time visual search engine that supports arbitrary keyword search combined with semantic analysis of topics;
- An enhanced real-time visual analysis system that enables users to explore and compare discussions and topics.

A preview of CompactMap can be seen in Figure 2.5.

Earthquake

Earthquake [9] is a visualization tool that uses Twitter posts regarding the earthquake which occurred on the East Coast of the United States (US) on August 23, 2011. It gathers information based on hashtags and displays the locations of different tweets in different time periods. It displays information as heat maps as well as graphs. An example of this tool can be seen in Figure 2.11.

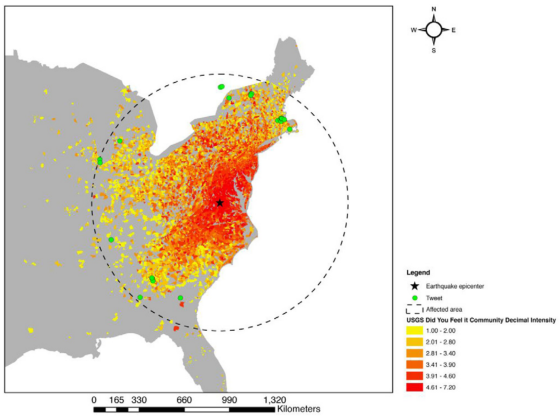


Figure 2.11: Earthquake visualization tool for 40 tweets [9].

A Software System for Data Mining with Twitter

A Software System for Data Mining with Twitter [39] is a visualization system designed to display Twitter information based on different queries. It gathers data posted from various geographic regions and then retrieves relevant tweets from that data using either spatial or textual queries. It enables the user to inspect retrieved tweets using a map based interface and graph/charts. Figures 2.13a and 2.13b give examples of analysis that can be done with this tool.

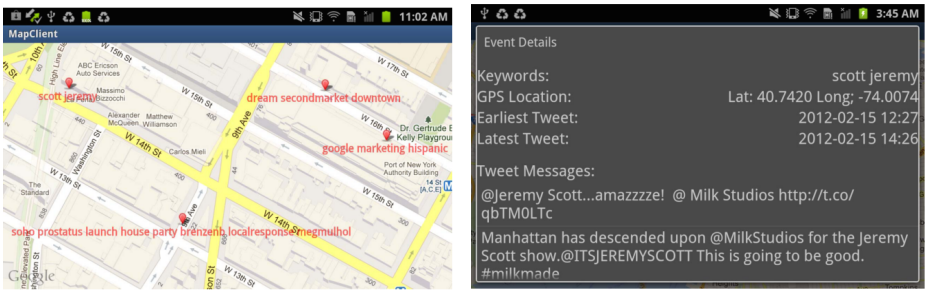
(a) [39] map visualization

(b) Distribution of tweets containing references to Smoking [39].

Figure 2.12: A Software System for Data Mining with Twitter.

EventRadar

EventRadar [31] is a novel local event detection method that analysis a seven day historic tweet data in order to improve precision. It uses maps visualization techniques in order to display the position of relevant tweets regarding certain events.



(a) A sample client using EventRadar [31]: a map showing local events.

(b) A sample client using EventRadar [31]: details of a selected local event.

Figure 2.13: EventRadar.

2.3 The TweeProfiles Project

TweeProfiles is a tool for multi-dimensional clustering and visualization of Twitter data. It analyzes geographical coordinates (spatial dimension), timestamp (temporal dimension), users (social dimension) and text (content dimension). This section gives a brief overview of Twitter as well as a simple description of TweeProfiles [17, 18] and its variants, Olhó-Passarinho [19] and of TweeProfiles2 [10].

2.3.1 Twitter

Twitter[4] is a social networking and microblogging service. It allows the users to broadcast short text messages with 140 characters in length, known as "tweets", to friends or "followers". Figure 2.14 shows an example of a tweet.

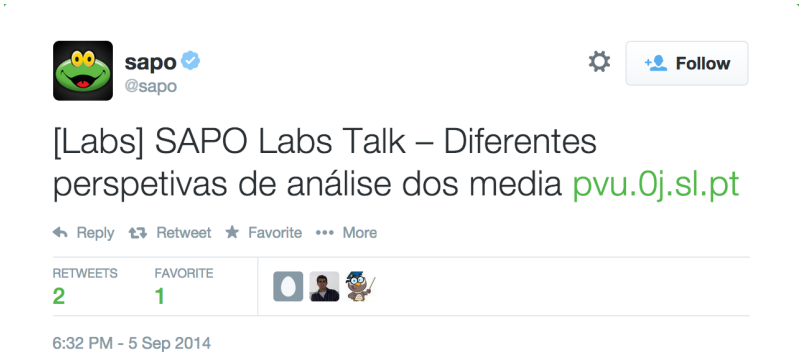


Figure 2.14: Example of a tweet.

Twitter is about discovering interesting people online and following their burst messages for as long as they are interesting. In Twitter, a social connection is made when a user follows or is followed by another one. A tweet is composed by different items and each may serve different purposes. These items can be for social interaction or to complete the information related to the message in question.

[4] https://twitter.com

- **Retweet (RT)** Share another user's tweet;
- **Mention (@ + username)** Identify a user in a tweet ;
- **Reply (@ + username)** Answer to a previous user tweet;
- **Hashtag (# + topic name)** Association of a keyword to a tweet;
- **Localization** User's geo-coordinates when sending the a tweet.

2.3.1.1 Twitter API

Twitter's information can be accessed by the use of two APIs: the REST API [40] and the Streaming API [41]. The first is request-based and requires an "oAuth" authentication while the latter uses events to provide information and requires either "oAuth" or HTTP basic authentication. The REST API allows access to information about the user, the timeline, friends and followers, direct messages geolocation, trends, and more. However, this API limits the number of requests allowed per user to either 15 or 180 tweets in a 15 minute window, depending on the method that is utilized. It is also limited by the availability of Twitter data and applications rate limits. The process for connection and data retrieval of the REST API can be seen in Figure 2.15.

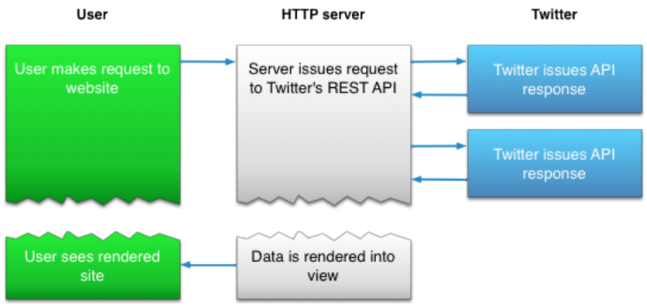


Figure 2.15: Interaction process of Twitter's REST API

The Streaming API uses real-time data, limiting the amount of data available according to the session's begin timestamp. This data is accessed through three types of streams: Public, User and Site. The first involves the public data available on this social network. In the others, filters are applied in order to only visualize tweets from certain users. Figure 2.16 shows the streaming API process for connection and data retrieval.

2.3.2 Máquina do Tempo

Máquina do Tempo (time machine) is an interactive tool that allows the navigation and exploration of news, from the last 25 years, of the Portuguese news agency (LUSA) file as well as the main websites for Portuguese news. With this system users can take a trip back in time, revisiting the most memorable events and personalities of Portuguese and International history in recent years,

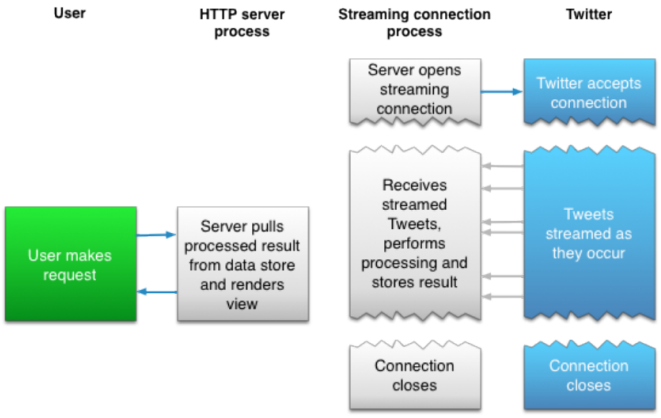


Figure 2.16: Interaction process of Twitter's streaming API

as they were portrayed by the news then published.

Figure 2.17: Results for Cristiano Ronaldo in *Máquina do Tempo*

Máquina do Tempo can be exploited in two ways: by choosing a personality or by choosing a specific date. When choosing a personality, a profile page in shown. It presents information collected from the news where that name is mentioned:

1. Generic personality profile data:
   - Name;
   - Last position or profession exercised;
   - Photographs (when available);
   - Global statistics, including the total number of stories in which the personality was present, as well as the total of their quotes and their connections.
2. A timeline from 1990 to the present day that allows you to see the evolution of the number of stories in which the personality in question was mentioned over time.

3. The most relevant content to the previously selected time period extracted from news

    - The list of news (with access to original content);
    - Quotes from the chosen personality that occur in those news;
    - The list of people that occur in those stories - the so-called "network connections" personality - which can be viewed on an interactive graphical format in the page itself, the so-called "network page"

When choosing a date without researching a specific personality, the user accesses a page centered in an interactive network. This network shows connections between personalities mentioned in the news in the defined time interval. Whenever two or more personalities are mentioned in the same news, their names are caught in the net and a link is established between them. Figure 2.18 gives an example of that network.

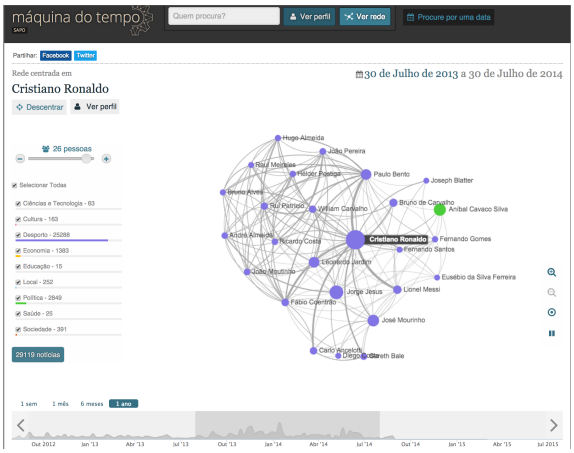


Figure 2.18: Network Connection for Cristiano Ronaldo in *Máquina do Tempo*

2.3.3 TweeProfiles Variants

The primary objectives of the TweeProfiles project are the development of a data mining platform to identify tweet profiles combining multiple types of data, the creation of a visualization tool to display the patterns found and to apply this tool on a case study in the Portuguese twittosphere [17]. It focuses on clustering tweets according to different dimensions: spatial, temporal, social and content. The spatial information is the location of the respective messages, temporal information is the date of the publication on Twitter, the social information is relative to the connection among users and the content is the text contained in each tweet. The goal of the original work was to verify if the combination of different dimensions for clustering would present interesting results and how they could be applied. The visualization tool was designed for a dynamic and intuitive use, enabling an interactive and comprehensible representation of the different profiles. TweeProfiles data collection and management is based on the [22] (previously known as TwitterEcho).
Olhó-passarinho [19] is an extension of TweeProfiles in which the main content dimension are

the images associated with tweets instead of text text. One of the main features of this extension is the combination of spatio-temporal dimensions (tweet's geographical coordinates and tweet's timestamps) with the content (images locally stored). The visual information is gathered with the aid of the VLFeat tool. The content is represented by a set of feature vectors of the images. The DBSCAN clustering algorithm is used with a combination of these three dimensions and the web application displays the resulting clusters. This has three main displays for the different dimensions used by.

TweeProfiles2 [10] is an extension of TweeProfiles [17, 18]. It addresses an important issue in the original version of the tool by replacing its batch clustering algorithm, with a method that is able to deal with stream data, DenStream. It transforms TweeProfiles into a real-time data analysis tool.

In the original version, the time dimension was represented by the time when the tweet is posted. However, since DenStream algorithm already applies a temporal decay factor to the data, the time dimension was represented by the hour in the day and the weekday of the post. To deal with the text data, the DenStream method was adapted in order to take into account all of the mentioned dimensions and not just the numerical ones.

2.3.3.1 TweeProfiles vs TweeProfiles2 vs Olhó-Passarinho

	TweeProfiles [17]	**TweeProfiles2 [10]**	**Olhó-Passarinho [19]**
Dimensions	Spatial, Temporal, Social, Content	Spatial, Temporal, Content	Spatial, Temporal, Content
Distance functions	Haversine, Time Difference, Geodesic, Cosine similarity	Haversine, Euclidean, Cosine similarity	Haversine, Euclidean
Distance normalization	Min-Max	Min-Max	Min-Max
Distance combination	On-demand	Online step: mixed; Offline step: on-demand	On-demand
Algorithm	DBSCAN	Micro-clustering: Hybrid DenStream; Macro-clustering: DBSCAN	DBSCAN
Data Structures	Dissimilarity matrixes	Hybrid micro-clusters	Dissimilarity matrixes
Clustering process	Offline	Online and Offline	Offline

Table 2.3: Differences between TweeProfiles, TweeProfiles2 and Olhó-Passarinho

All these variants of TweeProfiles focus on identifying tweet profiles through the analysis of different types of data. TweeProfiles used four dimensions for this task (Spatial, Temporal, Social, Content), while TweeProfiles2 and Olhó-Passarinho used only spatial, temporal and content. Olhó-Passarinho, being the only platform that includes images, is the only using a SIFT descriptor as the content. The bigger differentiating characteristic of TweeProfiles2, when compared to both

TweeProfiles and Olhó-Passarinho, is the ability to cluster in real-time. TweeProfiles2 structures the data in hybrid micro-clusters and uses a combination of a hybrid DenStream and DBSCAN, while TweeProfiles and Olhó-Passarinho apply the DBSCAN algorithm to dissimilarity matrices. Given that the dissimilarity matrices was a very costly and slow operation, TweeProfiles2 manages to obtain a faster and leaner clustering process. However, that efficiency comes with a price, as the clusters in TweeProfiles and Olhó-Passarinho effectively represent a group of tweets (the tweets that compose any given cluster are known), while the clusters in TweeProfiles2 represent a summary of the information of the tweets contained in the cluster (it is not possible to identify specific tweets compose a cluster).

2.3.4 TweeProfiles2

Figure 2.19 shows the architecture of TweeProfiles2. The data stream is pipelined to a back-end server where it feeds the micro-clustering algorithm, passing through several data pre-processing steps (see 3.1.1) beforehand. The resulting micro-clusters are then stored in a MySQL database, where they are accessible to the macro-clustering algorithm, whenever a clustering request arrives. The resulting clusters are then passed to the visualization module where they are to be displayed for analysis of the results [10].

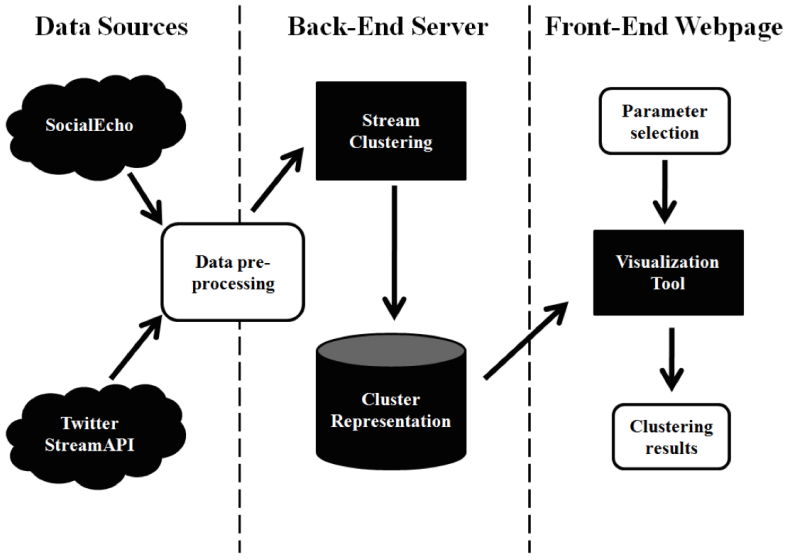


Figure 2.19: TweeProfiles2 high-level architecure

2.3.4.1 Operation

TweeProfiles2 was developed in Java, with a MySQL database used to store the micro-clustering results [10]. Several external libraries were used in order to complete some tasks along the entire KDD process. Table 2.4 describes those libraries.

Table 2.4: External libraries used in TweeProfiles2 [10]

Name	Description
MOA/Weka	MOA is a data stream mining framework which has a collection of several algorithms (including DenStream and DBSCAN) and evaluation methods. It works on top of WEKA, one of the most popular Java data mining frameworks
Lucene	Lucene is a text search engine library. It offers several text processing methods, including methods for stemming and stopword removal in texts, which are required in TweeProfiles2
org.json	Library that allows for the processing of JSON objects, which is the language used in the responses of the Twitter APIs
langdetect	Library used for detecting the language in which a text is written. Can detect up to 53 languages with a precision of 99%
MySQL connector	Driver for working with MySQL databases in Java

2.3.4.2 Illustrative Results

TweeProfiles2 offers three different sections for visualization of results and also includes controls for navigation and selection of certain parameters. One of the sections, as shown in Figure 2.20, consists of a world map where the red dots represent micro-clusters; the clusters are represented by a blue circular area, which is visible in the picture [10]. The clusters are plotted according to their center's geographical coordinates.

Figure 2.20: Spatial visualization

The temporal visualization consists of a x-y graph where the x axis represents the day of the week (Sunday through Saturday) and the y axis represents the hour. Clusters are represented by bubbles and are plotted according to their center's hour and weekday values, which defines the clusters center in the graph; the radius of the spheres represents the number of points (tweets) in that cluster. Figure 2.21 gives an example of this visualization type.

Figure 2.22 presents an example of the visualization of the content profile, namely text. This consist of a word cloud, where the size of each word represents its frequency. This means that the

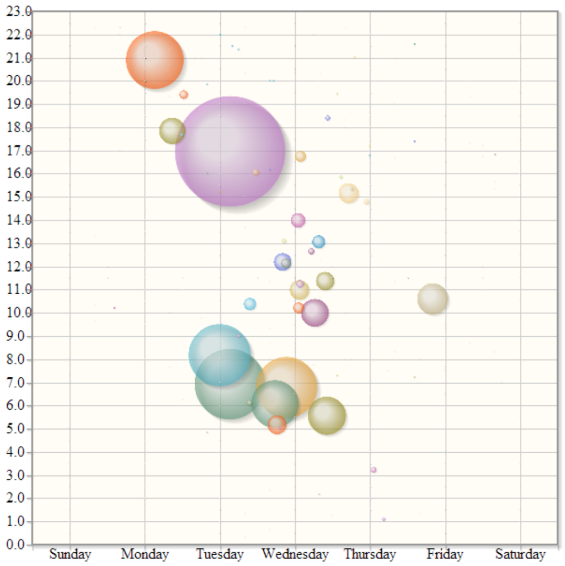


Figure 2.21: Temporal visualization

bigger the word is, the more times it appears in the texts of the tweets in the cluster [10].

Figure 2.22: Content visualization

The tool works by loading the (micro-)clustering results from a MySQL database and loading each cluster's information according to the visualization type [10].

Chapter 3

TweeProfiles3

In this Chapter it is explained the entire architecture of the developed tool, TweeProfiles3. In the following sections we present the results of the practitioner survey, provide an overview of the application and its architecture focusing on the integration between SocialBus and TweeProfiles2 and the different platforms from Sapo.

3.1 Introduction

As mentioned before, TweeProfiles2 was developed as an extension of TweeProfiles, allowing clustering in real time. The clustering is done over three dimensions: spatial, temporal and content dimension.
In order to obtain data from Twitter in real time, we integrated TweeProfiles2 in Social Bus. SocialBus [22] is a platform developed in Sapo Labs and it allows users to gather data from Twitter and store the desired information. This tool will be explained in more detail in Section 3.2.
Integrating these two tools is a big improvement to TweeProfiles2, due to the fact that we may restrict SocialBus to save only tweets with the required information for the task. Also, not only it allows TweeProfiles2 to be performed directly from the Twitter stream in real time, but also allows the information to be saved, in order to apply clustering to the same data. This is an important feature since this project is expected to be continued in the future with the objective of improving the algorithms used.
The key innovation presented here is the integration of linked visual-computational methods and a place-time-content conceptual framework in a working prototype grounded in both theory and practice informed by a structured survey of professionals.

3.2 System Architecture

Since one goal of this project was to integrate TweeProfiles2 in SocialBus, in order to define the architecture for this system it was necessary to understand the architecture behind these two systems. Section **??** introduced the architecture for TweeProfiles2 and it is represented in figure 2.19.

SocialBus is a research platform for extracting, storing and analysing the Twittosphere for R& D and journalistic purposes. Its architecture is presented in figure 3.1. SocialBus collects tweets in real-time using Twitter's Streaming API. These tweets are sent to a message broker (i.e., data format translator program) and processed over two components: stream processing and pre-processing. The resulting data is stored MongoDB. Figure 3.1 represents the complete system but only the most important steps were mentioned.

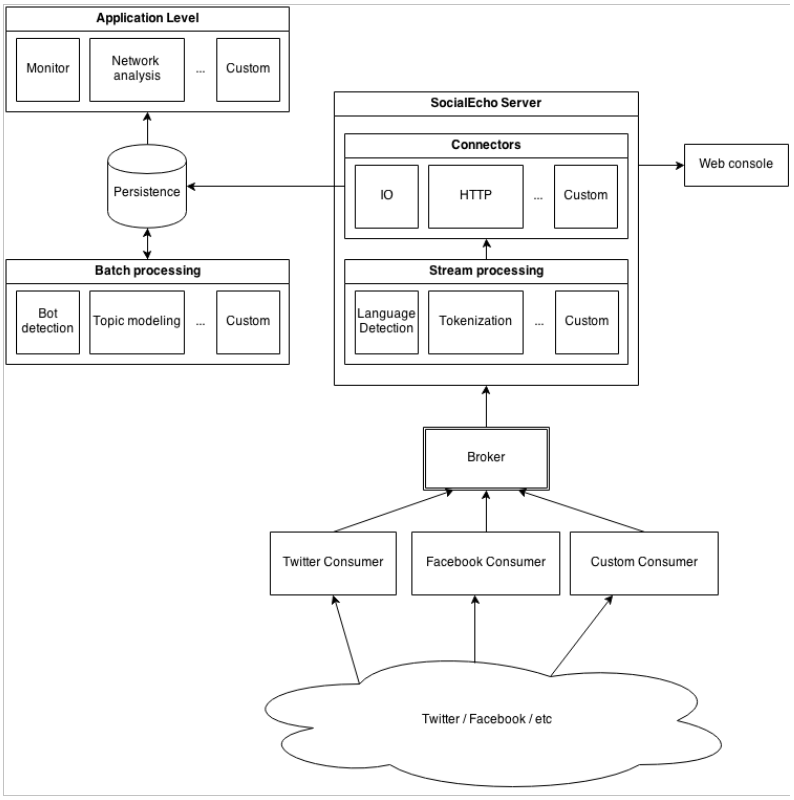


Figure 3.1: SocialBus high-level architecure.

TweeProfiles3 was defined by combining these two systems. We use SocialBus' extraction and pre-processing methods to gather data from Twitter and included TweeProfiles2 in the system so that when the data is stored, it automatically performs clustering. These results are then stored and mapped.

The use case diagram in figure 3.2 illustrates how users are expected to interact with the system. TweeProfiles3 must:

- Allow users to collect Twitter data posted from various geographic regions;
- Allow users to retrieve relevant tweets and clusters from the collected data using spatial queries;
- Allow users to retrieve relevant tweets and clusters from the collected data using textual queries;

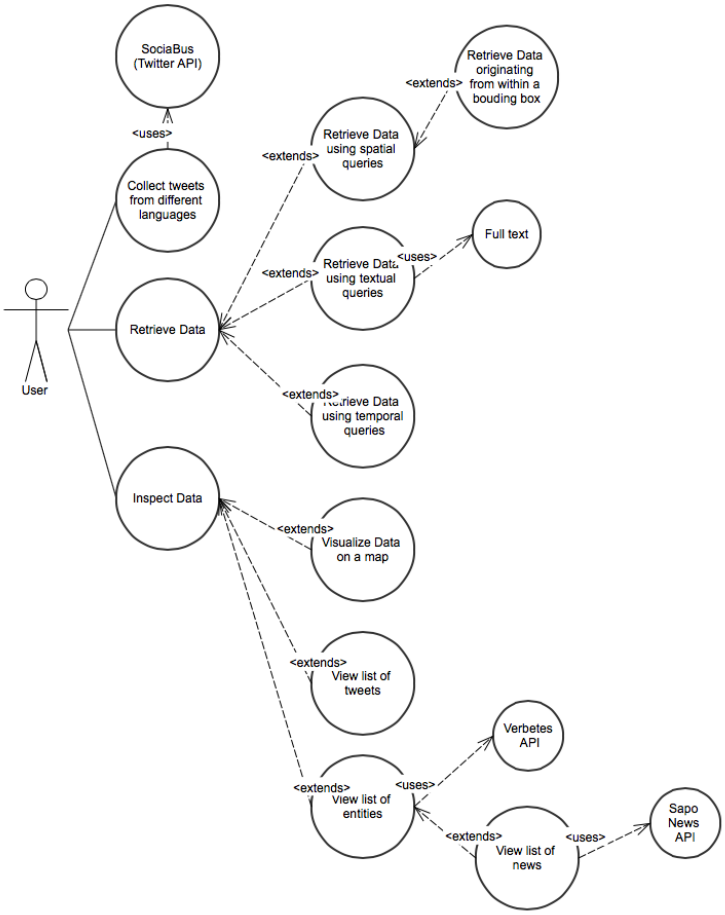


Figure 3.2: TweeProfiles3 use case diagram.

- Allow users to retrieve relevant tweets and clusters from the collected data using temporal queries;

- Allow users to inspect retrieved tweets and clusters using a map-based interface.

These requirements are fulfilled by several software modules that are detailed here.

Data Collection Module

The design of the data collection software module separates web-interface logic from the background processes for data collection and storage. This is completely done by SocialBus who uses the open-source Twitter4J library for connection to the Streaming API and stores received tweets in MongoDB. The web interface lets users monitor their progress. The architecture of this module was presented previously in figure 3.1.

Data Storage Module

The collected data is initially stored in MongoDB. The geo-located (latitude/longitude coordinates) and temporal information embedded in tweets must be represented in the correct format before search queries and distance measures can be applied on the data.
The data collected from the Streaming suffers a pre-processing done by SocialBus in order to convert date information to a usable format containing "year-month-day : hour-minute-second".

Data Retrieval and Clustering Module

Data Retrieval and Clustering are both integrated in the same module.
Design and implementation of the Data Retrieval module are straightforward after the conversion of Twitter data. MongoDB provides easy methods to retrieve stored data. Since not all the collected contains the required information for the clustering algorithms, in order to extend the spatial and temporal retrieval functionality offered to end users, some other pre-processing was required. This is will be detailed in Section 3.4.
As the data is retrieved from MongoDB it is passed to the clustering algorithms, DenStream and DBSCAN. DenStream is responsible for the creation of micro-cluster that are then used as input to DBSCNAN in order to create the macro-clusters. Both of these algorithms were implemented in Java.

Search Module

Full-text search is incorporated in our systems to assist end users with textual analysis, speed-up queries on a large data-set and to produce a broader set of search results for each keyword specified by the user. The system preforms a match of the desired word to both tweets and set of words in all clusters.
The system does not present an approach to deal with slang words and abbreviations commonly found in Twitter content. Since Twitter enforces a strict 140 character limit for each tweet, abbreviations and short forms of many words and phrases are widely used to preserve space. For example, the slang abbreviation 'IMHO' is used for shortening the phrase 'in my honest opinion'. As the majority of these abbreviations are not in dictionaries, they present a substantial challenge towards faithful analysis of Twitter content using natural language processing and data mining techniques.
The system also incorporates a full spatial search enabling the users the define a specific region to analyse, as well as a temporal search so that the user can define a certain weekday to display all tweets and clusters and a timeline to go back in time 7 days.

Data Mapping Module

The final module of TweeProfiles3 is used for data visualization. It is responsible for outputting the results of the clustering algorithms and of spatial, temporal and textual queries to end users.
The Google Maps and Leaflet [1] Javascript API are used to display retrieved tweets and resulting clusters on a map.

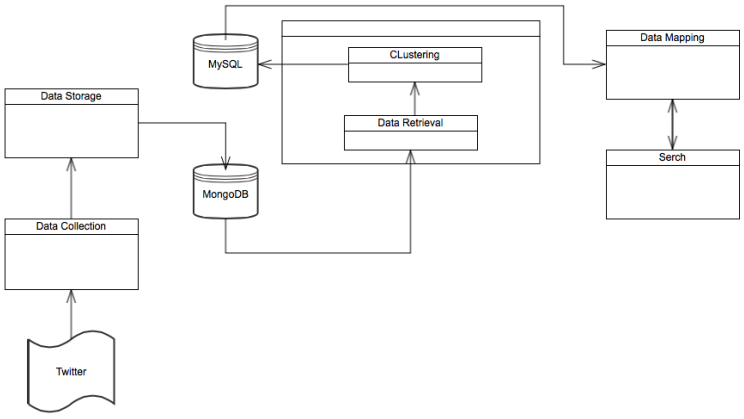


Figure 3.3: Modules from TweeProfiles3

3.2.1 Understanding User Needs

As an input to our tool design and development process, we completed a needs and interest study with two primary goals: (1) to develop a clear understanding of how media professionals use social media tools and (2) to draw out what journalist envision doing with social media geovisual analytic tools in the future. To address these objectives, a 32 question survey delivered online using Google Docs was created (this can be seen in Appendix A). To recruit participants were contacted people in the media field, namely people from FEUP, JornalismoPortoNet (JPN) and Publico P3. These interviews yielded 10 respondents for our survey.
This survey can be divided in three subgroups. One where we gathered information about the background of each participant, another where we asked about the use and knowledge of social media and the last where we asked participants to compare different applications similar to what we had planned.
The first question asked about work experience in media. All respondents were people with expertise and the majority had over 2 year of experience. The following Section reveals what we learned from the survey on our goals. The results are presented with percentages for questions that allowed multiple options to be selected, and by average subjective rating for questions using a 5-point Likert agree/disagree scale.

[1] http://leaftletjs.com

3.2.2 Use of Social Media and Geovisual Tools

To begin understating how social media is used by professionals, we asked participants to describe their use of social media tools in personal and professional environments. Facebook (100%), Twitter (100%) and Linkedin (80%) are most common for personal usage. In terms of professional usage, Facebook (100%) and Linkedin (60%) are most common. When it comes to journalism, Facebook (100%) and text messages (80%) are primarily used by the participants to communicate with other professionals, Facebook (100%) and Twitter (60%) to announce and gather information from the public, and Facebook (100%) and text messages (40%) to communicate with the public.
It was also asked the respondents what aspects of tools like Twitter they valued the most, and less, as a support platform of social media. The immediacy and interactivity were the aspects that participants described as the most important and the possibility that, in the chaos of such parallel information, followed by users, the journalistic message is lost, was the negative one. With that in mind, we tried to understand how useful these tools are, in regard to aspect such as instantaneity, interactivity, perpetuity, multimedia and hypertextuality. All participants agreed that Twitter is useful as a journalism tool to retrieve and share information due to instantaneity.
In the other major part of our survey, we tried to understand how the participants envisioned using geovisual tools that take advantage of social media data sources, namely Twitter. To begin, we asked participants to identify types of maps they would expect to see in new social media tools. They indicated that the locations mentioned in or relevant to the contributed information (100%) are more useful than the location of the person generating the information. Additionally, participants responded that they would find maps that show both types of geographic information at once to be better than just one.
It was then asked for the participants to identify certain tools and features expected in an interactive web-based application to incorporate social networks as a data source for media. Photos/video collections (100%), tables (60%), maps (60%), keyword clouds (60%) were among the most popular choices. After that we asked the participants to identify types of information they would expect to be able to analyze in the same application. Sequence of events (100%), people (100%), organizations (100%) and incidents (60%) were the most expected information types. In a final series of questions, we asked them to review a graphical mockup of the new TweeProfiles interface. First, we asked participants to describe how they might use such an application. Some of the answer we obtained were:

- *I'd choose the location, in order to see tweets from there, and most frequent words and related news.*

- *I'd use the information that was more prevalent and frequent, reconciling the map, the information of the clusters and tweets.*

After seeing the mockup, we demonstrated two other applications developed in Sapo Labs, the first TweeProfiles [2] and Retweet Pattern[3], and asked the participants to compared them with the interface shown.
Next, we asked the respondents what types of information or results they would expect to be able to share if they were to use an application like TweeProfiles3. This revealed that pre-formatted text reports (60%), printable maps (40%), a link that would launch the application with preloaded data (40%) and static screen captures (40%) were preferred.
The results of our survey suggest that those in social media, namely journalism, are actively using social media tools, that they expect to be able to explore multiple kinds of geographic information and anticipate to make use of that information. Our results provide a track for current and future designs of TweeProfiles that include functionalities for mapping tools, media integration and analytical reporting capabilities.

3.3 Visualization System

This Section introduces our application design process and initial system implementation and testing. First, we outline a base scenario, based on requirements and technologies, initial TweeProfiles3 functionality and present a platform analysis focused on key design choices.

3.3.1 Base Scenario

Our tool and interface design approach follows the results from our practitioner survey and understanding of users' needs.

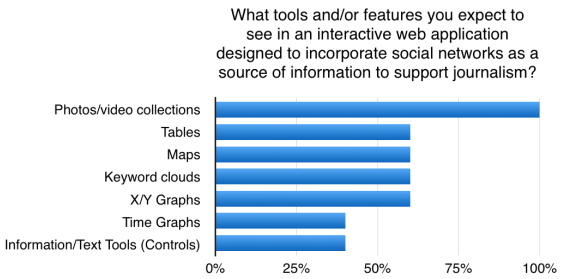


Figure 3.4: Response about tools/features expected in TweeProfiles3

Based on that knowledge, as well as the inquiry done for the state of the art, we defined a set of technologies to incorporate the majority of expected features and informations in a web application to support journalism.
Displaying a map with tweets' and clusters' information is a must, for an application of this type.

[2] http://reaction.fe.up.pt/tweeprofiles/tweeprofiles.html
[3] http://trodrigues37.github.io/RetweetPattern

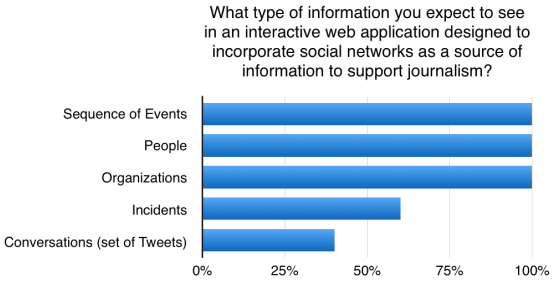


Figure 3.5: Response about informations expected in TweeProfiles3

We decide to use Google Maps and Leaflet, due to the fact that both are easy and accessible to use and contain all features required for the tasks planned. Since wordclouds and time graphs were among the expected features, we used two plug-ins from d3 [4] to display that information.

3.3.2 TweeProfiles3 Functionality

TweeProfiles3 exposes its functionality to end users through a php web application framework, designated Codeigniter [5]. It also provides a simple web based data collection and retrieval interface, which is shown in figure 3.6.

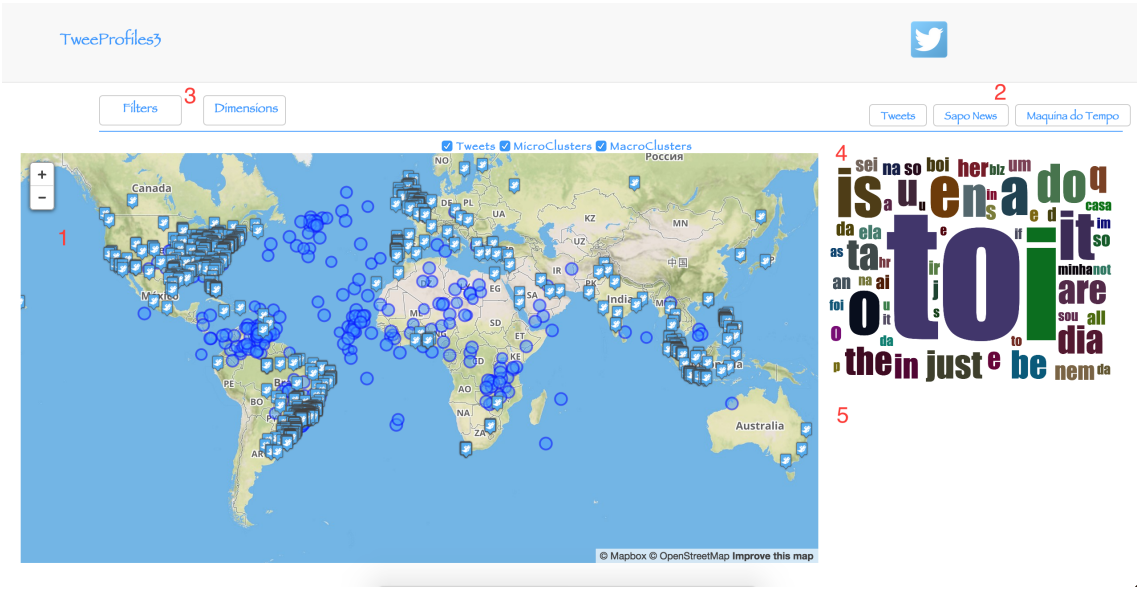


Figure 3.6: Screenshot of TweeProfiles3's web interface.

The system interface includes 5 core components: 1) Tweet and cluster map; 2) tweet, news and entities list; 3) display/dimension controls; 4) wordcloud; 5) time graphics. Each component is described in more detail below.

[4] http://d3js.org
[5] http://www.codeigniter.com

Tweet and Cluster map: Following the logic outlined previously, the map supports simultaneous tweets and clustering distribution overview. In one platform, a heatmap provides the overview for tweets and the clusters are displayed using simple markers. In a second platform, tweets' position is displayed using markers and clusters with circles (figure 3.6). Here we can get detailed information such as the number of words in the cluster, the creation time and position.

Tweet, News and Entities list: The tweet list depicts the 1000 newest tweets for any query. The entity list displays any personality found in the processed tweets and the news list display the 20 more recent news for those entities from Sapo.

Display/Dimension controls: The Display controls support query filtering with region selection, time range sliders (either choosing weekday or going back in time) and a minimalistic interface to specify terms for text-based queries. These controls also include a filter to specify the desired dimensions.

Figure 3.7: Screenshot of TweeProfiles3's filters interface.

Wordcloud: The wordcloud displays the most frequent words in all clusters. The size of the words is equivalent to its frequency in tweets. The more it appears, the bigger the word's size is.

Time graphics: These graphics display the hour and weekday of the creation of all clusters. The size of the cluster corresponds to the number of points it contains.

All widgets are related. When selecting an element in the map, all other widgets (wordcloud and time graphs) change in relation to that. Besides that, if a macro-cluster was selected, only micro-clusters related to that macro will be shown in the map. This can be seen in Section 3.4.3.

3.4 System Implementation

In the following Section we present the implementation of TweeProfiles3.

3.4.1 SocialBus meets TweeProfiles2

As mentioned previously, one goal for this work was to fully integrate TweeProfiles2 in SocialBus. Both of these platforms were developed entirely using the Java language for all the algorithm and processes with MySQL and MongoDB database to store information. These two databases were chosen due to the fact that MongoDB provides easy methods to access and process the data stored, so it made easier to save and analyse tweets that came from the stream and MySQL has simple

methods to connect and retrieve data from a web interface, so it was used to store results from the clustering. From the external libraries used in TweeProfiles2 listed in Table 2.4, only MOA/Weka and MySQL connector were used, since SocialBus already enabled all the rest. MOA was the library chosen for our task, because it is a proven stream data mining framework where DenStream is already implemented.
The Twitter consumer inside SocialBus pipelines the data stream to a back-end server where it builds the micro-clustering algorithm, passing through multiple pre-processing steps (Section 3.4.2). These micro-clusters are then saved in the MySQL database and used by the macro-clustering algorithm. The resulting clusters are stored and both are then passed to the visualization module, where they will be displayed for analysis.

3.4.2 Data Processing

The data used by the clustering algorithms is extracted from the tweets. However, the JSON response from the Twitter API provides over 30 fields per tweet, some of them containing nested objects. Most of that information is useless for our needs. Therefore, the first pre-processing is applied as the data is retrieved from MongoDB, where only some of the fields are selected.


```
{
tweetid:527145570964881408
lat:-52.414141
lon:-29.74042
date:2014-10-28T17:11:12.000Z
username:Vivizinhaa6
tweet:minha turma  um inferno
}
```


Figure 3.8: Example of tweet after the first pre-process.

After this step, some extra operations are required before passing the data to the micro-clustering algorithm. The first process involves extracting the hour and weekday from the date string. The range value for these attributes are 0-23, for the time (hour) and 1-7 for the days of the week (from Sunday to Saturday).
After the date is processed, we need to address the content of the tweet by executing the following tasks:

1. Detect the language of the text;
2. Remove any URLs it contains;
3. Remove all punctuations from the text;
4. Tokenize the text;

All the steps are pretty straightforward and self-explanatory.

3.4.3 Visualization

After obtaining the clustering it is important to display them in an intuitive and simple way. All state of art representations presented in Section 2.2, together with the results of our survey, were considered to provide the best visualization system.

The chosen strategy for the system assumes a spatio-temporal representation as the basis with the other dimension overlaid on top of those, as it can be seen in figure 3.6. This decision was based on the fact that time and space dimensions are the most intuitive and easily representable and interpretable.

As previously mentioned, the visualization system was made through a php web application framework and different JavaScript widgets. Each time the platform is loaded (site.php), the map is automatically centered at $(0,0)$ and a pre-established zoom level is used.

The representation of the spatial dimension is made through an interactive map on which is plotted referenced tweets and clusters. Tweets are represented by either a heatmap, with the density of tweets for area, or by a simple 'marker' object with fixed size and an icon representative of Twitter. This enables to discern the position of all tweets.

On top of this, clusters are represented by circles. Each circle has its own size and location according to the position of tweets on the map. The center of the cluster is obtained by the average of the latitude and longitude.

This representation approach can be seen in figure 3.9. Like [17] [18], it was based on the methodology used by [33] but with the same two differences: the colors to represent each cluster and its points are not the same and the shape of the clusters is a circle instead of an ellipse. This is due to the fact that we could not link tweets to micro-clusters, since the algorithm implemented in TweeProfiles2 did not provide that option. We decided to plot all markers and circles with the same color.

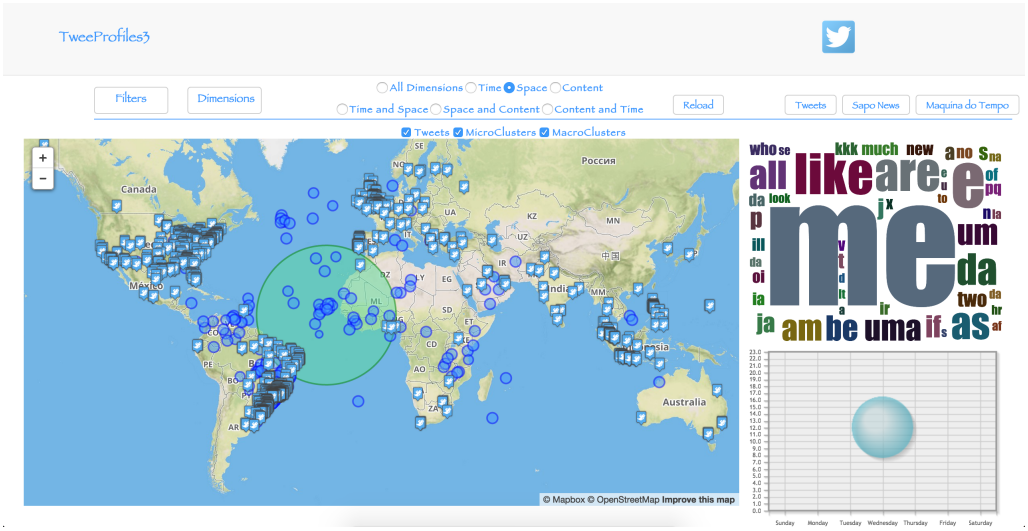


Figure 3.9: Base representation of TweeProfiles3

One disadvantage of this type of representation is the overlapping markers and circles when zoomed out. Even though some of these are associated, there are some whose size and location are the same (or almost the same) which makes only the ones on top accessible. The only solution for this problem is zooming in a specific region and/or using the available filters.
The temporal representation, as shown in figure 3.10, consists of a $x-y$ graph where the x represents the day of the week and the y represents the hour. Each cluster is represented by a bubble and is plotted according to their center's hour and weekday. The radius of the bubble is the number of tweets in the cluster. The temporal visualization is also based on a timeline using horizontal bars to characterize each cluster. The length of the bar is given by the earliest and latest timestamps of that cluster. An example of this can be seen in figure 3.11.

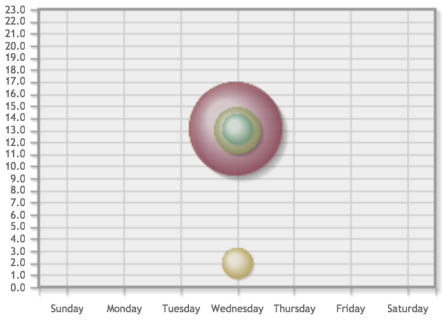


Figure 3.10: Temporal visualization - $x-y$ graph.

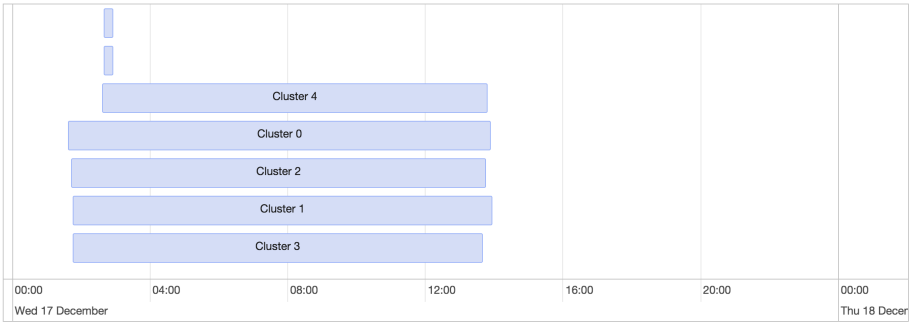


Figure 3.11: temporal visualization - timeline.

The remaining dimension, content dimension, is presented next to the temporal one, as seen in the base representation of TweeProfiles3 (figure 3.9). It consists of a wordcloud containing the most frequent words where the size of the word is equivalent to its frequency. The bigger the word is, the more times it appeared in the tweets. Clicking any word in the cloud performs a search for news relative to that word in *Sapo.pt*.
All dimensions are presented in different parts of the platform, but they can easily be related. Clicking on a tweet/cluster causes the map to adjust to that location and to display an 'Info Window', when using Google Maps, or a 'popup' object, when using Leaflet. Clicking on a cluster also shows the wordcloud and time graphic for that specific one. If that cluster is a macro-cluster

it also shows all micro-clusters contained in that specific macro. Figure 3.12 demonstrates that information.

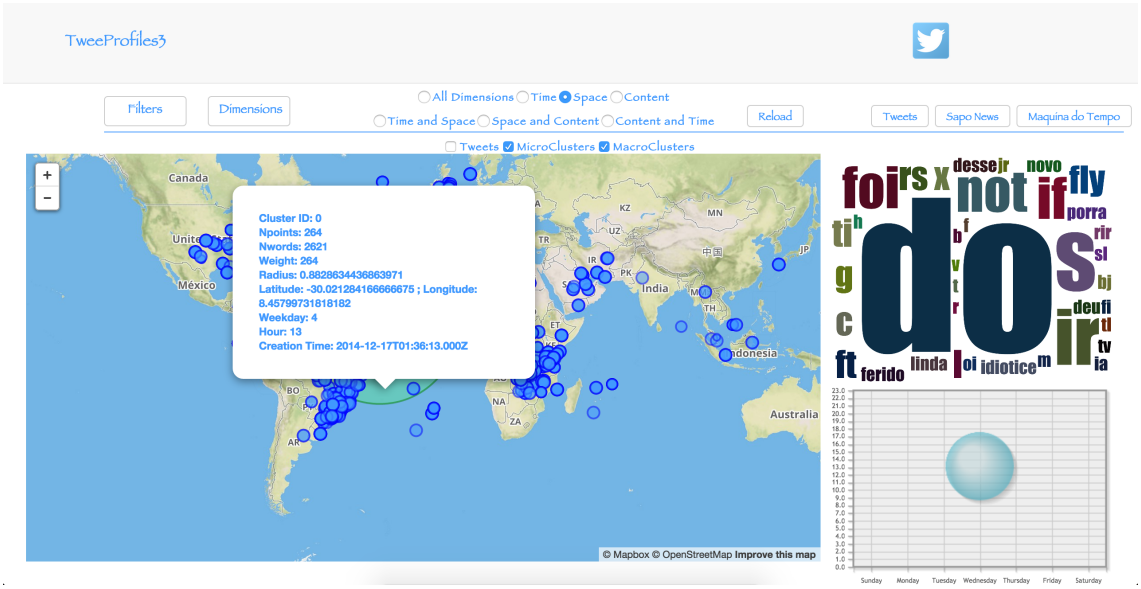


Figure 3.12: Popup with information regarding one cluster.

Due to the complexity of the mined data and the purpose of this platform, interaction is a decisive factor. It must not only provide the ability to set new parameters but also to enable users the navigation in the results. Parameters can be set using multiple filters and selections available. All filters and dimension selection use Asynchronous JavaScript and XML. The purpose of this feature is to enable users a more detailed analysis of the data.

3.4.4 Sapo Platforms

After all data goes through both clustering algorithms and is sent to the visualization module, an other API is used, *getPersonalities* from *Verbetes*. This is used in order to detect entities from the Tweets and connect those entities to the *Máquina do Tempo*. In summary, a list of names and IDs is obtained from the API and a direct match is done from the tweets to that list. If a name is found the system inserts that entity in an array and presents it in the visualization, enabling an URL to both *Máquina do Tempo* and Sapo News. Figure 3.13 gives an example of some entities found in tweets.

Besides presenting the entities found, the system also presents a list of recent news for each person. It uses a second API, provided by Sapo, to obtain that list. The tittle, where it was published and a URL to the original news is given to the user, as it can be seen in figure 3.14.

Figure 3.13: Example of entities in the testing dataset.

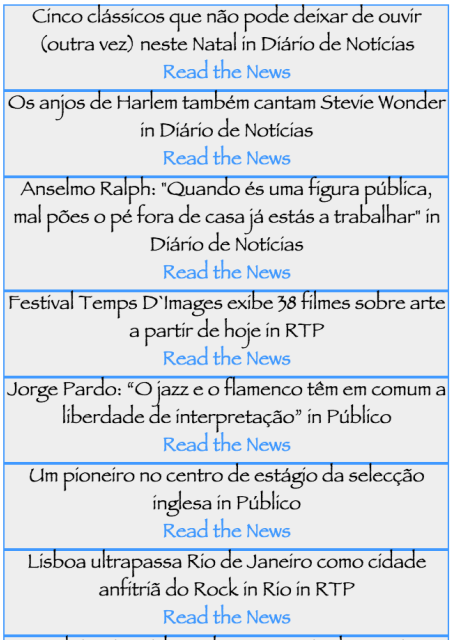


Figure 3.14: Example of news obtained from the list of entities.

Chapter 4

Results

In this Chapter we present some of the results obtained in the tests carried out. The system software was tested for a large scale data over a period of 30 days (starting mid December). The data collection, for testing purposes, was focused on tweets labelled in Portuguese, English, French and Spanish. Figure 4.1 shows the volume in the dataset used for the tests and figure 4.2 the position of the tweets. All these tweets were retrieved without using any specific query method, besides the language filtering.

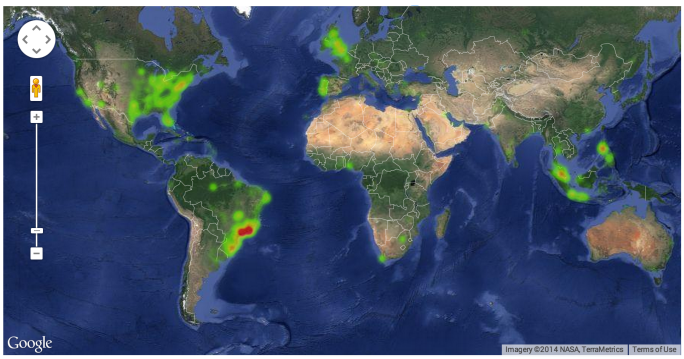


Figure 4.1: Volume and spatial distribution of tweets in the testing dataset.

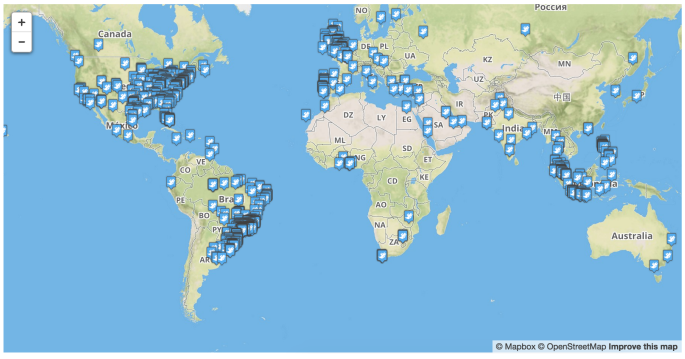


Figure 4.2: Position of the tweets in the testing dataset.

4.1 Clustering

To access the algorithm, we performed a series of testing rounds with different component combination (for DBSCAN). Table 4.1 shows all the tests done to data gathered.

Combination	**Spatial Dimension**	**Temporal Dimension**	**Content Dimension**
1	100%	100%	100%
2	100%	0%	0%
3	0%	100%	0%
4	0%	0%	100%
5	100%	100%	0%
6	0%	100%	100%
7	100%	0%	100%

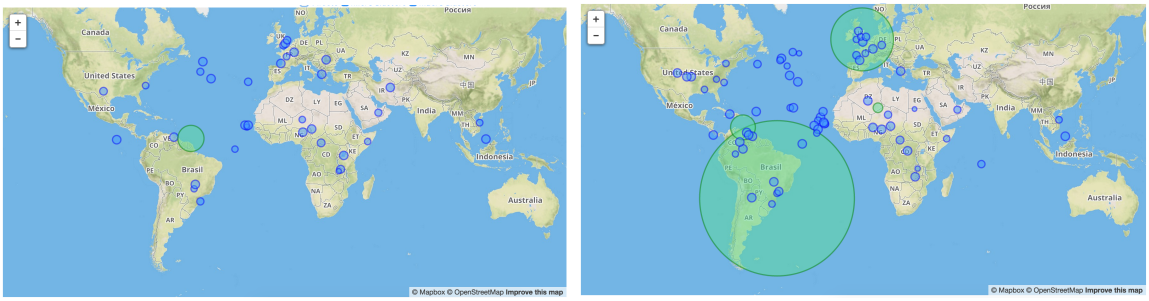
Table 4.1: Set of tests performed.

Since the distance function weighting was not performed correctly in TweeProfiles2, we decided not to include that option, turning all distance functions into a binary variable. Besides that, HybridDenStream has different parameters that can be altered to produce different results for the clustering.

Name	**Abbrev**	**Description**	**Min Value**	**Max Value**	**Default Value**
Epsilon	eps	Defines the minimum radius of a HMC	0	1	0.1
MinPoints	mp	Defines the minimum number of points in the ε-neighbourhood to create a HMC (also used as μ parameter)	1	∞	2
InitPoints	ip	Number of points for initialization	50	∞	1000
μ	μ	Used in the PMC/OMC restriction	1	∞	1
Beta	β	Used in the PMC/OMC restriction	0	1	0.2
Lambda	λ	Used in the time decay function; affects the decay rate of the stream	0	1	0.25
Processing Speed	s	Defines the number of instances (tweets) per time unit	1	∞	100

Table 4.2: HybridDenStream parameters.

DBSCAN also has two input parameters, epsilon and MinPoints. These have the same meaning as the one mentioned above, but are used for the macro-clustering process. In order to obtain



(a) Time point 1 (b) Time point 2

Figure 4.3: Spatial Clustering results.

the results shown below, as the main goal of this project was to extend TweeProfiles2 and not improve the robustness of the clustering algorithm, all these parameters were set to default value, besides InitPoints that was set to minimum. DenStream and DBSCAN's parameters (eps and MinPoints) were set to 0.4 and 1 (for DenStream) and 0.4 and 2 (for DBSCAN).
All the results demonstrated below were obtained with these values.

Spatial Dimension

The results of this testing round were obtained with the second combination of table 4.1, meaning only the spatial distances were taken into account. Figure 4.3 shows the clusters obtained for this dimension. In this representation, each blue circle represents a micro-cluster and all green circles, the macro-clusters. All micro and macro clusters' radius are calculated based on its radius and the number of points. Each time point is a moment in the simulation where we stored the clustering results. It is useful to see the evolution of the clusters and in this case, since only spatial information is being processed, how the macro-clusters appear in a large concentration of micro-clusters.
The resulting clusters can be deemed adequate, due to the fact that with these languages being processed, most of the tweets will be coming from Europe and South America.

Temporal Dimension

When it comes to temporal dimension, some problems were faced regarding the clustering. Since we used the default parameters for both algorithms and the micro-clusters are created taking all dimensions into account, DBSCAN creates multiple macro-cluster with very similar timestamps.
Even thought we faced this problem, it can clearly be seen in figure 4.4 that the system creates one cluster including the majority of points, when the dataset is from a single day only. Figure 4.5 shows that big difference between the resulting clusters.

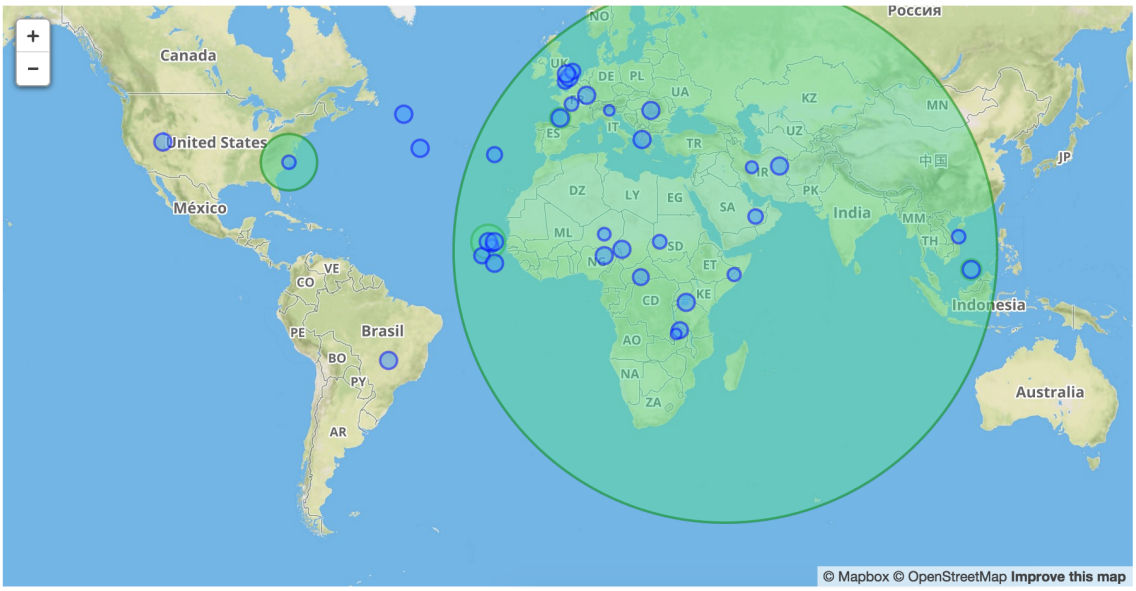
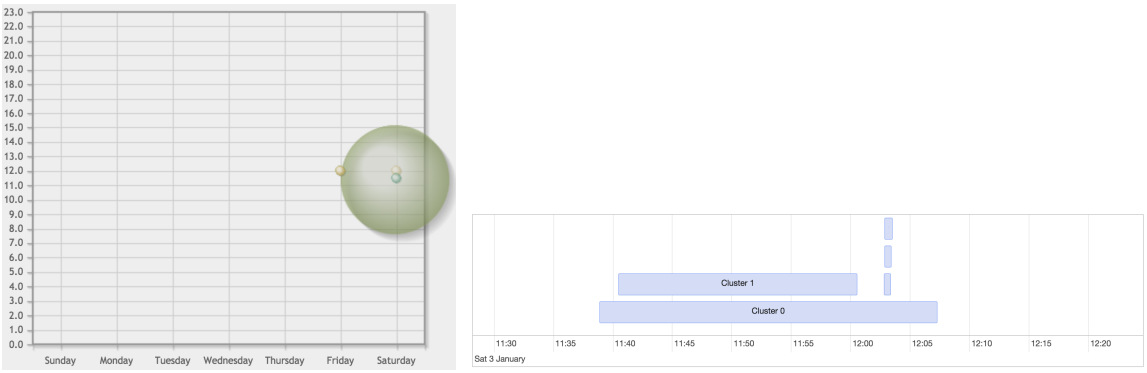


Figure 4.4: Temporal Clustering results.

A big difference can be seen when including the spatial dimension to this combination. Figure 4.6 presents the results for this clustering.

Content Dimension

For this round of tests, it was only taken into account the content of the tweet (combination 4 from table 4.1). The results are not as satisfactory as we would want, due to the fact that compared texts are small, the computed similarity is also small.
Even thought the ε values are high, the number of resulting clusters for this dimension is small. However, if it increased even more the quality of the results may decrease, because the algorithm will be merging text instances that have little similarity. This was a problem also faced in TweeP-



(a) Time point 1 (b) Time point 2

Figure 4.5: Temporal Clustering results.

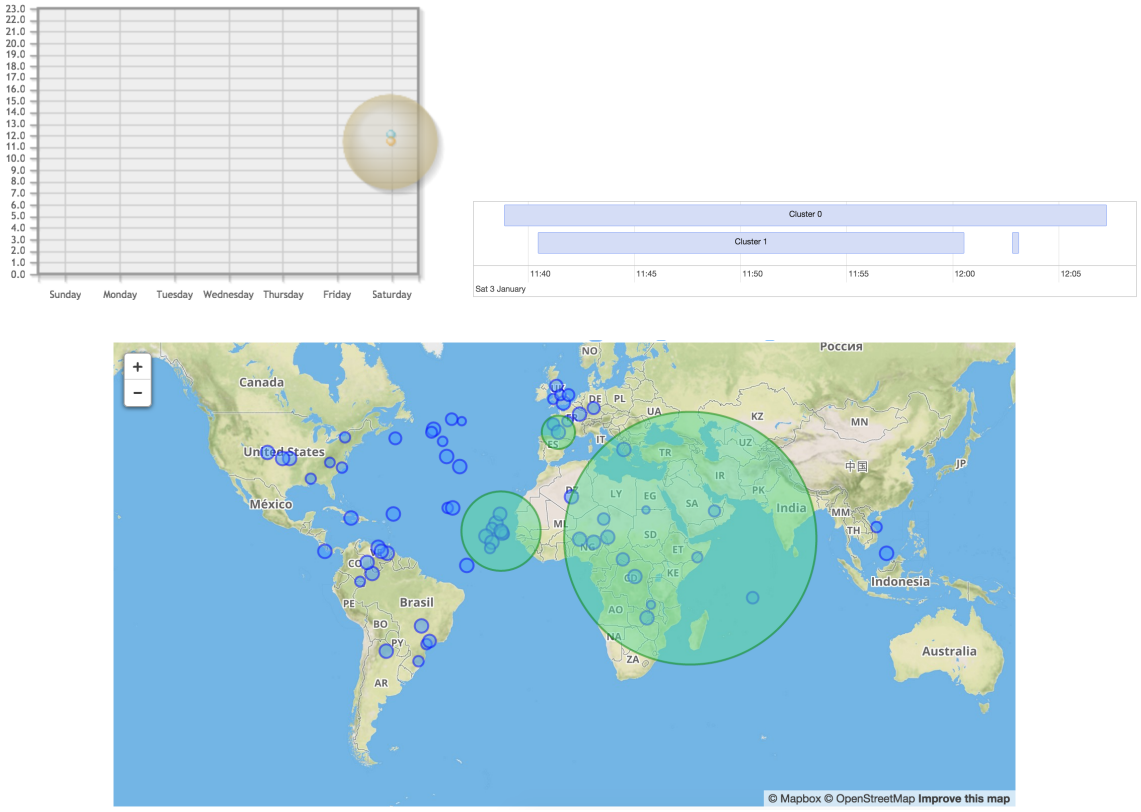


Figure 4.6: Temporal and Spatial Clustering results.

rofiles2.

Figures 4.7, 4.8 and 4.9 give the results of a test of the content dimension. It presents the evolution of a cluster over different periods of time. It shows, for each time point, the cluster and the most frequent words formed at that moment. The size of the words is proportional to their frequency relative to others. It can be seen that, even thought they differ in size, the most common words remain consistent.

Like the previous dimension, a difference can be seen when including spatial distances to the clustering. Figure 4.10 shows how the cluster changes when only using the content dimension. Figures 4.6 and 4.10 can also be compared to 4.3 where a big difference can also be seen.
Even though the differences are hard to evaluate, in some cases it can be seen how the outcome reflects when including other dimensions to the clustering. All results presented demonstrate some interesting patterns.

The results obtained were satisfactory in the sense that it made possible to understand better how the algorithm reacts to SocialBus' data extraction. It provides a fully integrated real-time clustering system with a state of the art visualization platform. With the clustering being done in real-time but not being optimized, the results obtained here are not as good as the results in the latest publication of TweeProfiles [18]. Figures 4.11 and 4.12 present two examples obtained.

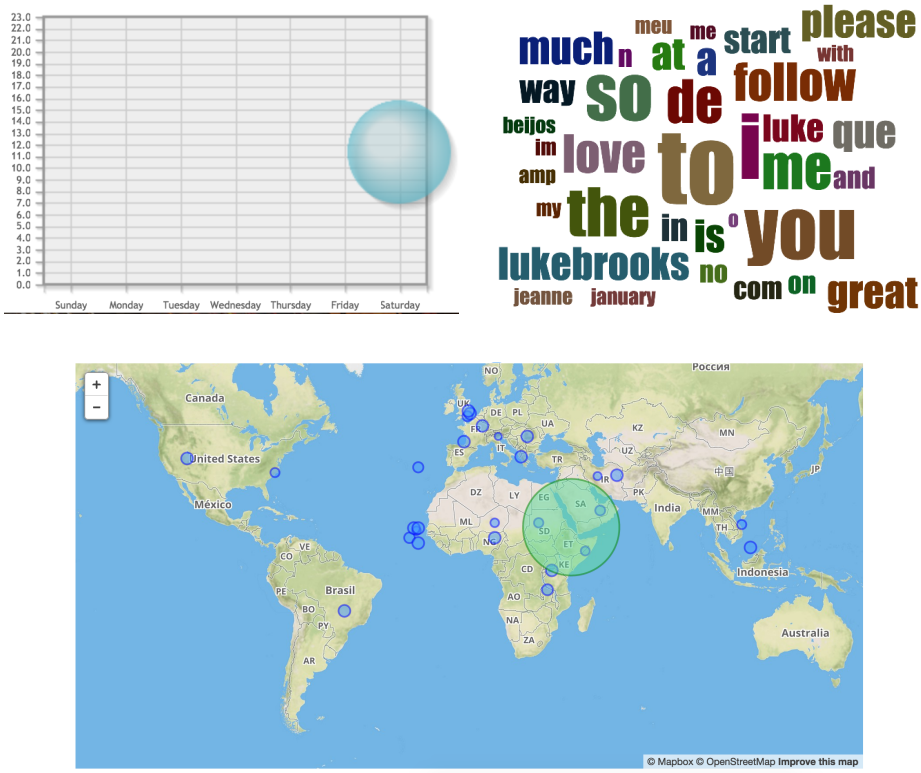


Figure 4.7: Content Clustering results.

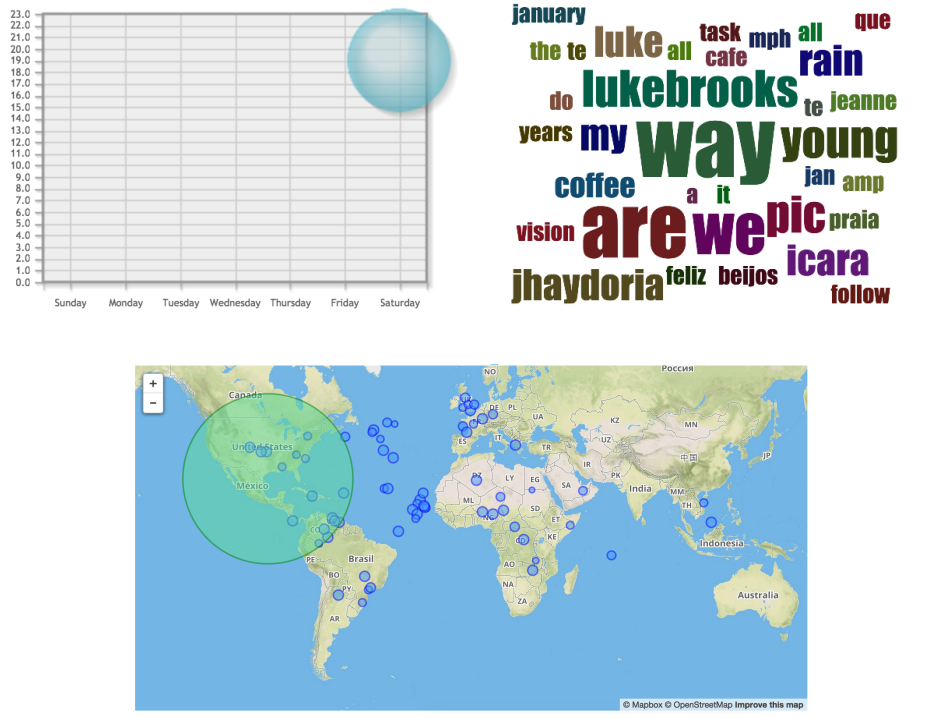


Figure 4.8: Content Clustering results.

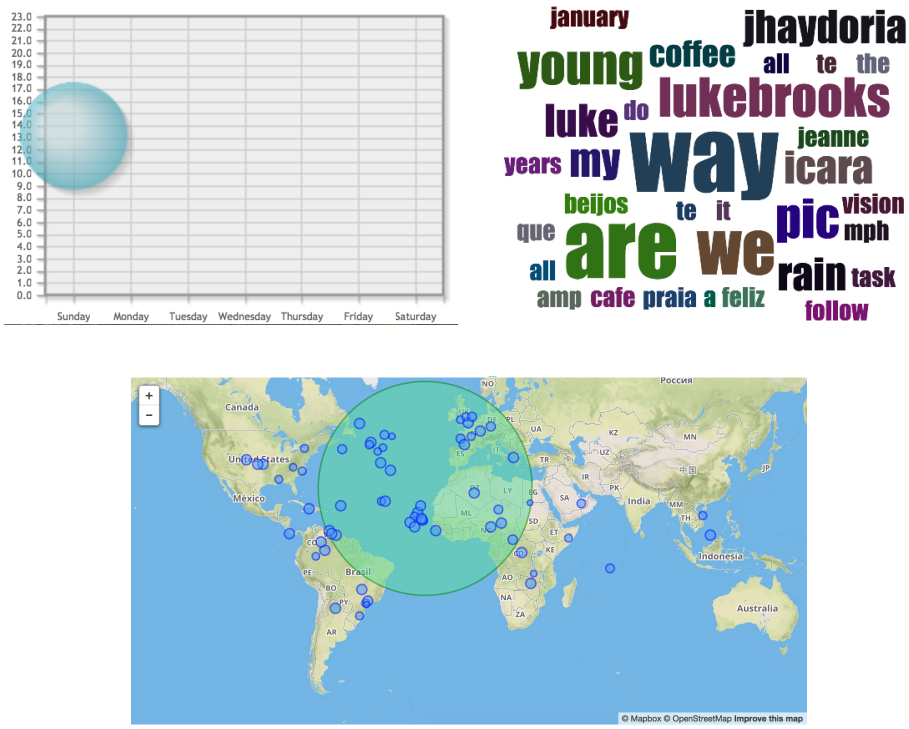


Figure 4.9: Content Clustering results.

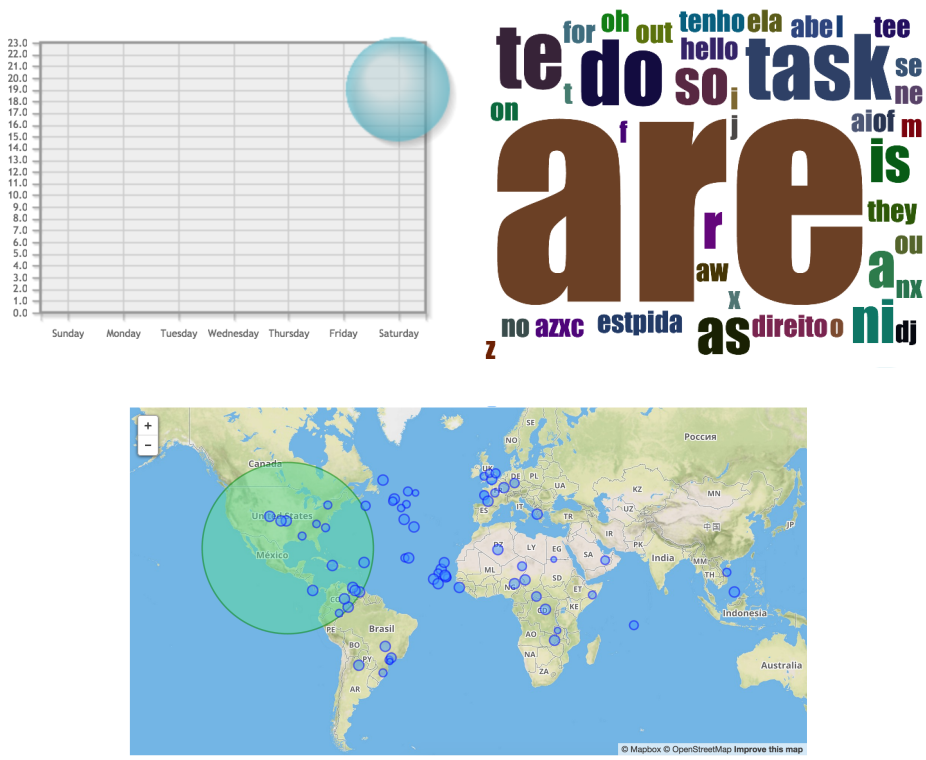


Figure 4.10: Content and Spatial Clustering results.

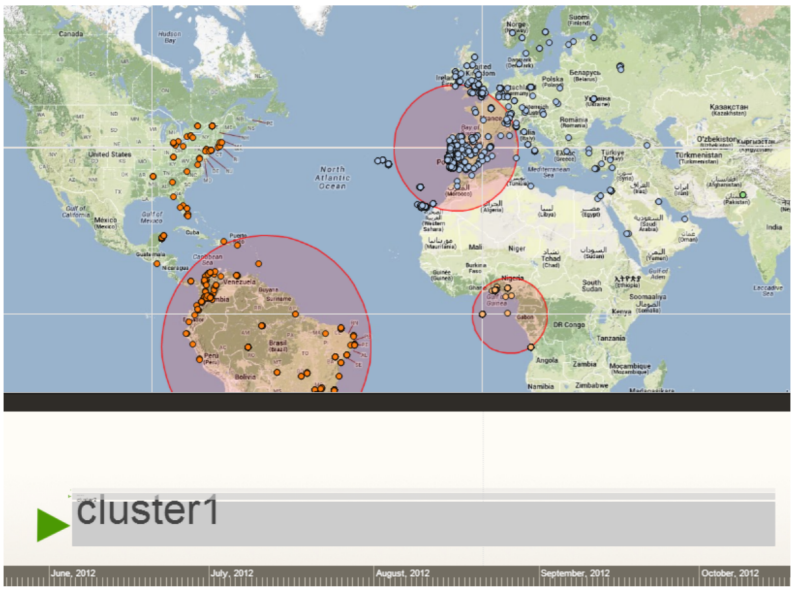


Figure 4.11: TweeProfiles spatial clustering results.

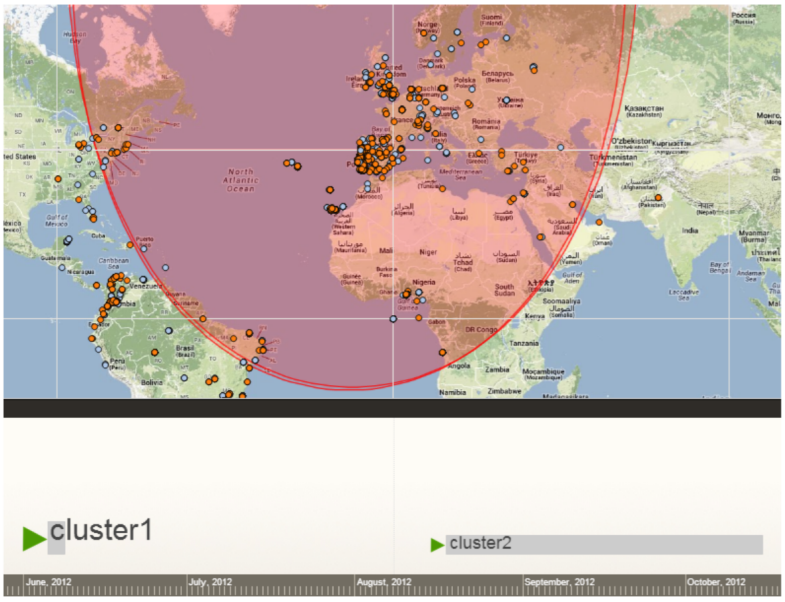


Figure 4.12: TweeProfiles temporal clustering results.

When comparing these example to the spatial (figure 4.3) and temporal (figure 4.5) clustering results from TweeProfiles3, it is clear that TweeProfiles provides the best performance. However, it does not enable a real-time data analysis, which is a big advantage of our system.

4.2 Usability Tests

As a method to understand if TweeProfiles3 met the desired features and design of end users, we completed a second set of tests (this can be seen in Appendix B). Here we made the system

available and asked the participants to use it to extract information from tweets, news and clusters, evaluating the time and detail of the result. These tests were made with people from JPN, but unfortunately only 3 were available.

While the participants were using TweeProfiles3, we were able to understand that they easily and effectively extracted information from both tweets and all news features. The ability to apply different filters to the data was highly appreciated and used, being the content filtering the one that got the most attention. All data shown on the map seemed to be useful for the tasks and the platform displaying the position and details of the tweets, and not just the density, was the one participants preferred, the reason being that qualitative information is as important as quantitative. An example given for the usage of tweet's details was the terrorist attack on Charlie Hebdo. They said that TweeProfiles3 would be perfect to support an article regarding what happened, since they could not only gather information from the news, but also from what people were saying, giving the possibility to insert that in the article.
Regarding the same topic, Charlie Hebdo, one aspect considered missing were hashtags and images. These two features are removed from the analysis and since the biggest trend from the attack included images and hashtags, a lot of information was lost. Improving the algorithm to use this would be a big step to TweeProfiles3.
One other aspect considered important by the journalists was the integration of Sapo platforms in TweeProfiles3. Even though the news list was more analysed than *Máquina do Tempo*, both were acknowledge as a plus.
One feature that didn't get as much attention from those using our tool, was the different combination of dimensions we included. Spatial and temporal were the dimensions that participants most used and analysed. This may be due to the fact that these presented better results and because what interests journalists the most is what people are saying regarding a topic of interest and where they do so.

With the system used, participants were asked to complete a form, for us to understand how they felt about the implemented features and the general use of TweeProfiles3. This survey was created based on [42] and people were able to evaluate, from 1-5, different features and aspects of the system (1 being that they disagreed with the statement and 5 that they fully agreed).
At first people evaluated how simple the usage of the system was. All answers point to the fact that TweeProfiles3 is a simple system to use (4, on our scale). This was our biggest concern when designing the application, to create a simple system with all the desired information. Our respondents also agreed (with the same evaluation) that they were able to complete work effectively using TweeProfiles3.
Regarding the usability of our system, one participant evaluated with a 4, and one of them with a 3, on our 1-5 scale, how easy it was to understand.
When it comes to the information we present, all respondents answered with 5 on how clear the information was organized on the screen, but regarding how easy it was to find the information

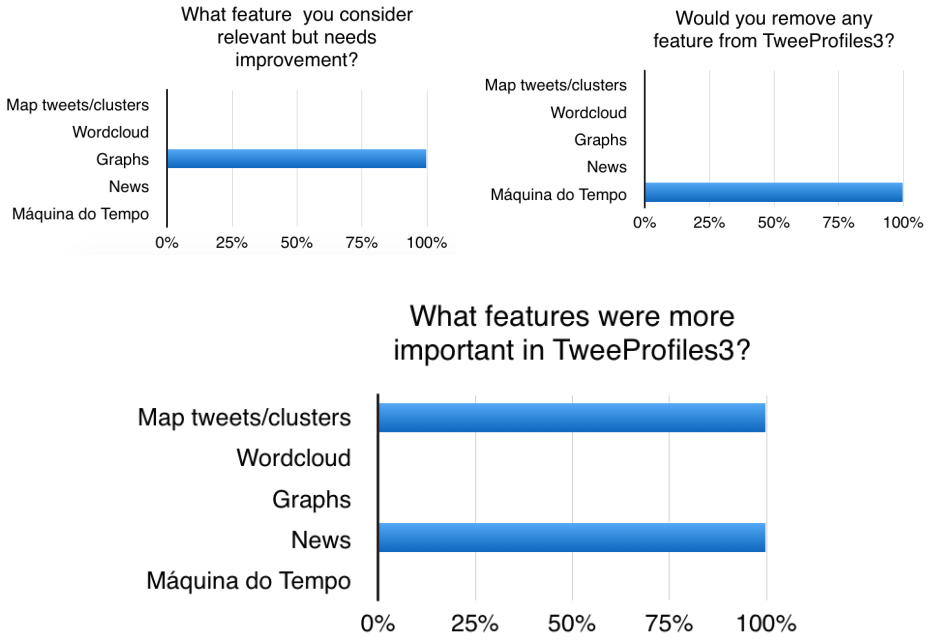


Figure 4.13: Survey questions regarding features from TweeProfiles3.

they were searching, one person evaluated with a 3, while others still evaluated with a 5 (all on previously mentioned scale).
One other question all participants came to agreement was regarding the implemented features. All said that the most important were the map with all tweets/clusters and the news and one person answered that they would remove the list of entities from *Máquina do Tempo*, since only known people could be searched. That same person said that a feature which required improvement were the time graphs and that the words in the wordcloud would need more work, in order to be more detailed. Other participants left these questions in blank.

When asking if they considered the integration of the platform from Sapo, all participants answered positively and also, all agreed that the news list was more relevant than *Máquina do Tempo*. An aspect needed and desired to improve the usage of TweeProfiles3 is the possibility of defining topics of interest, such as football, politics, entertainment... This would enable journalist to visualize data regarding only the theme they would be interested in, which would bring more detail to the extracted information. Evaluating all these features, asking if our platform included everything they were expecting, the answers were 3 and 4 on our scale.
When finally asking the participants if they were satisfied with the outcome of TweeProfiles3, the answers were positive, all being 4, which means our system proved to be to the liking of all people who tested it.

Even though there are some improvements to be made and some features to be added, and the number of journalists we had to test our application was fairly low, it was easy to understand

how all participants considered TweeProfiles3 faster, more efficient and detailed than any other traditional method for gathering information. Platforms like this are seen as useful and necessary (which was answered on our survey with a 4, from all respondents) for the journalism community, and overall, TweeProfiles3 went in the right way, satisfying all participants.

Chapter 5

Conclusions

In this chapter we present a summary of the project and achieved goals, a discussion of some decisions and results and some limitations of both the clustering and visualization. We finish this with some suggestions for future work.

5.1 Summary

The goals of this project were to develop an integration of a real-time clustering algorithm and SocialBus, to create a visualization system capable of displaying the profiles found in all different dimensions, as well as different news and entities from Sapo and to take the first in evaluating the design tool as a real-world application. The purpose of creating a system with all this information is to support the user in journalism while gathering detailed data.
In order to accomplish these goals, we develop a system with the following stages: data extraction, data pre-processing, clustering and visualization.

Data Extraction: We were to obtain a stream of tweets connecting TweePofiles to SocialBus. It enables the use of real-time data or a static data set to perform clustering.

Data Processing: Before using the extracted data in both algorithms, each tweet needs to be filtered and pre-processed so we can obtain the desired features we want the clustering to be based on.

Clustering: Since the main goal of this dissertation was not to improve the stream algorithm, we decided to follow the previous work. So, the clustering was performed based on both algorithms used in TweeProfiles2, DenStream and DBSCAN. These perform a multi-dimensional clustering over the tweets' data, considering its content and spatio-temporal location. Most of the parameters from both algorithms were set to default, which influenced the resulting clusters, but

did not invalidate them.

Visualization: The visualization tool made use of different visual patterns associated with all dimensions. Specifically, a map with information regarding both tweets and clusters, a timeline and a graph with timestamp details and a wordcloud with the most common words in all clusters. All these widgets were implemented using various JavaScript libraries. It enables a simultaneous representation of the same information in different dimensions and to interact for a deeper and more detailed analysis of the data.

User studies: In order to complete this project, two surveys were made with journalists. The first was made to gather user needs for a platform such as TweeProfiles3. Here, we were able understand the usage of social networking tools by media professionals and to draw out what journalists envision doing with social media geovisual analytic tools. The second one, a usability test, was made so people could use TweeProfiles3 to extract information from tweets, clusters and news, and for us to understand the efficiency of the platform.

All the objectives set for this dissertation where met, besides the inclusion of the social dimension. It was an improvement over TweeProfiles2, not only when it comes to the visualization tool, but also to the data gathering methods. It generates clustering results in real-time without the need to locally store the data used by the algorithms. However, in the following section we will be discussing some of the decisions made and results obtained and suggest some alternatives.

5.2 Discussion

In the developed system, there are some theoretical decisions that influenced the performance and results. Regarding the data mining process, aspect such as:

- **Combinations:** We decided to use all combinations possible to access the algorithm. Since the weighting of the dimension was not implemented, we had to use a binary variable to include or remove a dimension to the clustering. However, this was only performed in DBSCAN and not in both algorithms. This does not lead to the best results, because, for example, in the combination where we only take the spatial dimension into account, all other dimensions were used when creating the micro-clusters. This can be fixed by implementing a method to weight the dimension in both DenStream and DBSCAN.
- **Clustering Parameters:** Setting all clustering parameters to default is arguable. The problem is that each dimension dissimilarity distribution, even though they have the same scale after the normalization, does not have the same shape. One value does not represent the same in all dimensions. An increase of 10% on space may represent a change of continent and in time it may only mean one day apart. Since the main goal of this dissertation was not to improve or evaluate the clustering algorithm, our decision was that these values had to be

automatically chosen, due to the large number of clusterings being executed. Although we don't have a complete understanding of the impact this decision had on the results obtained, a deeper analysis should be made on this matter.

- **Evaluation:** The evaluation of the clustering, in this project, was very simple: we manually selected some interesting cases to discuss and performed a usability test with the main users for a platform such as this. The main reason for this is that clusterings are difficult to evaluate as the task is unsupervised. This has been a problem following TweeProfiles since [17] [18]. A possible solution could be to use some general measures of clustering quality.

When it comes to the visualization tool, there were also some debatable decisions, at a more technological level.

- **Database:** In order to store the results of the mining process, a MySQL database was used. This was necessary due to the JavaScript libraries and the php framework used. Even though this was an effective solution for our problem, it is not the perfect solution as the system response time is sometimes affected by the large number of requests. Another possible solution is to use MongoDB to store both tweets and the clusters and access the data through a RESTful web service.

- **Visualization Widgets:** The widgets used in this tool followed the approaches in other state of the art projects. As far as our knowledge goes, our choices reveal appropriate and useful. But many other widgets or filters can still be incorporated in the system leading to a more detailed analysis of the data.

- **Overall usage:** Due to the computational cost of the clustering, all parameters of the algorithms are hidden from the users. This can be made accessible in possible future versions, along with the dimension weights, with a more stable and powerful algorithm.

5.3 Future Work

Although we feel the goals for this dissertation were met, there are some aspects that can still be improved. We believe that a project with the magnitude that TweeProfiles has reached can always be better, whether by changing the complexity of the data mining process or by extending implemented features in the visualization tool. Some of the most important aspects to be improved are:

- **Data:** Some additional filter could be applied to the data chosen to our data mining process. This would enable the retrieving of results from a niche of observations. Besides the language filter being used, another possible option would be to extract data based on regions. One could filter the data for Portugal and apply the data mining process in a more focused manner.

- **Clustering Process:** The algorithm used for this project is not the most effective when dealing with small batches of data over a long period of time. This is also due to the fact that some of parameters were also not thoroughly tested and could be tuned, if the time is spent doing the necessary testing. This would presumably lead to better clustering results. The distance combination also needs some improvement. It is necessary to implement a method of weighting all distances before the clustering, so the macro-clusters are not compromised by other distances. Using that, the number of combinations are extended and it would be possible to investigate the best compromise between the differences among the clusters obtained and the computational costs of the process.

- **Evaluation:** A complete evaluation is a great step for this project. Not only does it validate the approaches chosen but also it would be possible to tune the parameters required for clustering.

- **Social Distance:** As mentioned before, the social dimension was not included in TweeProfiles3. It raised a problem that has still not been resolved. It is an important issue that should be improved, because it can give users an deeper analysis about the social interactions in Twitter.

- **Visualization System:** The ability to choose topics had already been discussed when developing the system and when doing the second test, that need was confirmed by all participants. This would be a great step for the future of TweeProfiles3. The need to improve graphs and the detail in the wordcloud was another aspect we were able to understand from the usability test.
  More tests can be made in the future, since more people will be available in JPN. This was already discussed and should be scheduled for the second semester.

Finally, it is important to mention that a fourth version of TweeProfiles is schedule to happen in the near future. This dissertation will focus on improving the algorithm used by the data mining process, which will lead to new and, hopefully, better clustering results.

Appendix A

Appendix A - Questions made in the first Survey

- Qual é o seu nome?
- Qual é a sua idade?
- Quais as suas habilitações académicas?
- Qual é a sua função profissional atualmente?
- Alguma vez trabalhou numa área relacionada com o jornalismo/comunicação social?
- Se sim, quanto tempo?
- Quais as redes sociais que conhece/usa?
  - Facebook
  - Twitter
  - Linkedin
  - Myspace
  - Outra
- Usa alguma rede social profissionalmente?
  - Facebook
  - Twitter
  - Linkedin
  - Myspace
  - Outra
- Alguma vez usou uma rede social para comunicação com profissionais (relativamente a jornalismo/comunicação social)?

  - Facebook / Myspace
  - Twitter / Microblogging
  - Mensagens de Texto
  - Não usei
  - Outra

- Alguma vez usou uma rede social para divulgar informações ao público (relativamente a jornalismo/comunicação social)?
  - Facebook / Myspace
  - Twitter / Microblogging
  - Mensagens de Texto
  - Não usei
  - Outra

- Alguma vez usou uma rede social para reunir informações do público (relativamente a jornalismo/comunicação social)?
  - Facebook / Myspace
  - Twitter / Microblogging
  - Mensagens de Texto
  - Não usei
  - Outra

- Quais são os aspetos que mais valoriza em cada uma das ferramentas que conhece, como uma ferramenta de apoio ao jornalismo/comunicação social?

- Quais são os aspetos que menos valoriza em cada uma das ferramentas que conhece, como uma ferramenta de apoio ao jornalismo/comunicação social?

- Ferramentas de microblogging como o Twitter são úteis, como ferramenta de jornalismo/-comunicação social, para a recolha de informação
  - Concordo plenamente
  - Concordo
  - Não concordo, nem discordo
  - Discordo
  - Discordo plenamente

- Ferramentas de microblogging como o Twitter são úteis, como ferramenta de jornalismo/-comunicação social, devido à instantaneidade

  - Concordo plenamente
  - Concordo
  - Não concordo, nem discordo
  - Discordo
  - Discordo plenamente

- Ferramentas de microblogging como o Twitter são úteis, como ferramenta de jornalismo/-comunicação social, devido à interactividade
  - Concordo plenamente
  - Concordo
  - Não concordo, nem discordo
  - Discordo
  - Discordo plenamente

- Ferramentas de microblogging como o Twitter são úteis, como ferramenta de jornalismo/-comunicação social, devido à perenidade
  - Concordo plenamente
  - Concordo
  - Não concordo, nem discordo
  - Discordo
  - Discordo plenamente

- Ferramentas de microblogging como o Twitter são úteis, como ferramenta de jornalismo/-comunicação social, devido à multimediação
  - Concordo plenamente
  - Concordo
  - Não concordo, nem discordo
  - Discordo
  - Discordo plenamente

- Ferramentas de microblogging como o Twitter são úteis, como ferramenta de jornalismo/-comunicação social, devido à hipertextualidade
  - Concordo plenamente
  - Concordo
  - Não concordo, nem discordo
  - Discordo

  - Discordo plenamente

- Que tipos de informação geográfica associada a redes sociais considera relevante para uma ferramenta de jornalismo/comunicação social?
  - Localização da pessoa que está a gerar a informação
  - Localizações que são mencionadas ou relevantes para a notícia (locais mencionados no Tweet sobre um evento, por exemplo)
  - Outra

- Que tipos de mapas associados a redes sociais considera úteis para o jornalismo/comunicação social?
  - Mapas que mostrem a localização da pessoa a partilhar a informação
  - Mapas que mostrem localizações mencionadas ou relevantes para a notícia (locais mencionados no Tweet sobre um evento, por exemplo)
  - Mapas que mostrem as duas anteriores
  - Outra

- Que ferramentas e/ou características espera ver numa aplicação web interativa desenhada para incorporar redes sociais como fonte de informação para apoio ao jornalismo/comunicação social?
  - Tabelas
  - Mapas
  - Gráficos de Tempo
  - Chave / Tag Clouds
  - X / Y gráficos
  - Processamento de texto e ferramentas de informação Ferramentas de Clustering
  - Ferramentas para Animação
  - Modelos preditivos
  - Gráficos de rede sociais
  - Coleções de Fotos/Vídeos
  - Outra

- Que tipo de informação espera ver numa aplicação web interativa desenhada para incorporar redes sociais como fonte de informação para apoio ao jornalismo/comunicação social?
  - Incidentes
  - Pessoas

  - Organizações
  - Sentimentos
  - Sequência de Eventos (timeline)
  - Resumos de Conversação
  - Informação Geográfica
  - Outra

- Interface Por favor reveja o design da interfaces que se segue.

  Representa um esboço de design que será desenvolvido para uma aplicação web interativa que iria apoiar o pessoal de jornalismo/comunicação social que precise de extrair informações ou partilhar notícias em redes sociais. Este esboço usa o Twitter como fonte de informação. Este design apresenta múltiplas componentes; (1) um mapa interativo que oferece suporte básico e mostra os locais dos tweets assim como os clusters identificados pela plataforma; (2) sistema de visualização de clustering, capaz de mostrar os tweets que o compõem, assim como palavras chaves de cada perfil; (3) uma ferramenta para definição dos pesos atribuídos a cada uma das dimensões; (4) uma cloud de tags que sumariam palavras comuns nos tweets; (5) lista de tweets baseado nas queries feitas com a dimensão do tempo, espaço e conteúdo.

  Se fosse usar esta aplicação (ou uma semelhante) como suporte ao jornalismo/comunicação social, como usaria?

- Compare o esboço com as seguintes aplicações (Demo 1)
  TweeProfiles - http://youtube.com/watch?v=tpEZULbDHY4
  RetweetPattern - http://youtube.com/watch?v=69JwdAmqqgc

- Mudaria a resposta à pergunta anterior (relativamente ao esboço), tendo em conta as aplicações demonstradas?

- Que informações/resultados espera ser capaz de partilhar através desta aplicação (ou uma semelhante)?

  - Relatórios préformatados
  - Link que inicia a aplicação & data
  - Mapas para a impressão
  - Capturas de ecrã estáticas
  - Pequenos clips de vídeo
  - Outra

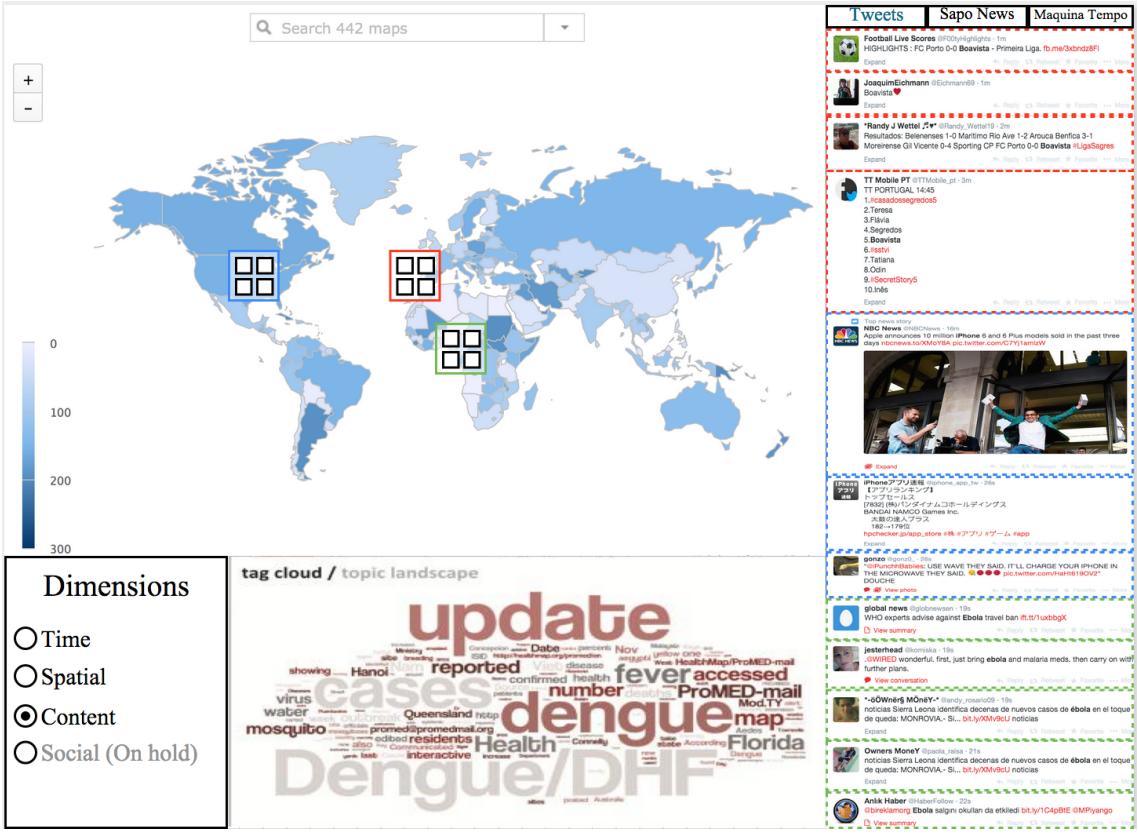


Figure A.1: TweeProfiles3 mockup.

Appendix B

Appendix B - Questions made in the second Survey

- Qual é o seu nome?
- Qual é a sua idade?
- Quais as suas habilitações académicas?
- Qual é a sua função profissional atualmente?
- Alguma vez trabalhou numa área relacionada com o jornalismo/comunicação social?
- Se sim, quanto tempo?
- É um sistema simples
  - 1
  - 2
  - 3
  - 4
  - 5
- Consigo completar o meu trabalho eficazmente com este sistema
  - 1
  - 2
  - 3
  - 4
  - 5
- Foi fácil de entender a usabilidade do sistema

  - 1
  - 2
  - 3
  - 4
  - 5

- Foi fácil encontrar a informação que procurava
  - 1
  - 2
  - 3
  - 4
  - 5

- A organização da informação no ecrã é clara
  - 1
  - 2
  - 3
  - 4
  - 5

- O sistema contém as funções e ferramentas que estava à espera
  - 1
  - 2
  - 3
  - 4
  - 5

- Estou satisfeito com o TweProfiles3
  - 1
  - 2
  - 3
  - 4
  - 5

- Que características do TP3 achou mais importantes?

  - Mapa com tweets/clusters
  - Wordcloud
  - Gráficos
  - Notícias
  - Máquina do Tempo

- Retirava alguma ferramenta da aplicação? Qual?
  - Mapa com tweets/clusters
  - Wordcloud
  - Gráficos
  - Notícias
  - Máquina do Tempo

- Alguma característica que considera relevante mas que necessita de melhoria?
  - Mapa com tweets/clusters
  - Wordcloud
  - Gráficos
  - Notícias
  - Máquina do Tempo

- Considera as ferramentas de ligação ao Sapo uma mais valia?
  - Sim
  - Não

- Qual foi mais relevante?
  - Notícias
  - Máquina do Tempo

- Considera a informação (palavras) contida nos clusters relevante?
  - Sim
  - Não
  - Precisa de mais detalhe

- Considera a possibilidade de escolha de temas (futebol, política, entretenimento...) uma caraterística importante para o futuro?
  - Sim

  - Não

- Compare a recolha de dados no TP3 com um método tradicional. Qual é mais rápido, eficiente ou detalhado?
- Vê necessidade/utilidade numa ferramenta como o TweeProfiles3?
  - Sim
  - Não
- No geral, estou satisfeito com o resultado final
  - 1
  - 2
  - 3
  - 4
  - 5

B.1 Planning: Initial v Final

Our initial planning for the dissertation didn't suffer major changes through out the whole development process. Some problems were encountered regarding TweeProfiles2 and SocialBus, leading to some delays and to turn the completion of the social dimension into a secondary objective, due to lack of time. We also added a second set of interviews, in order to fully understand the usage of the developed system. Overall, the schedule was met.

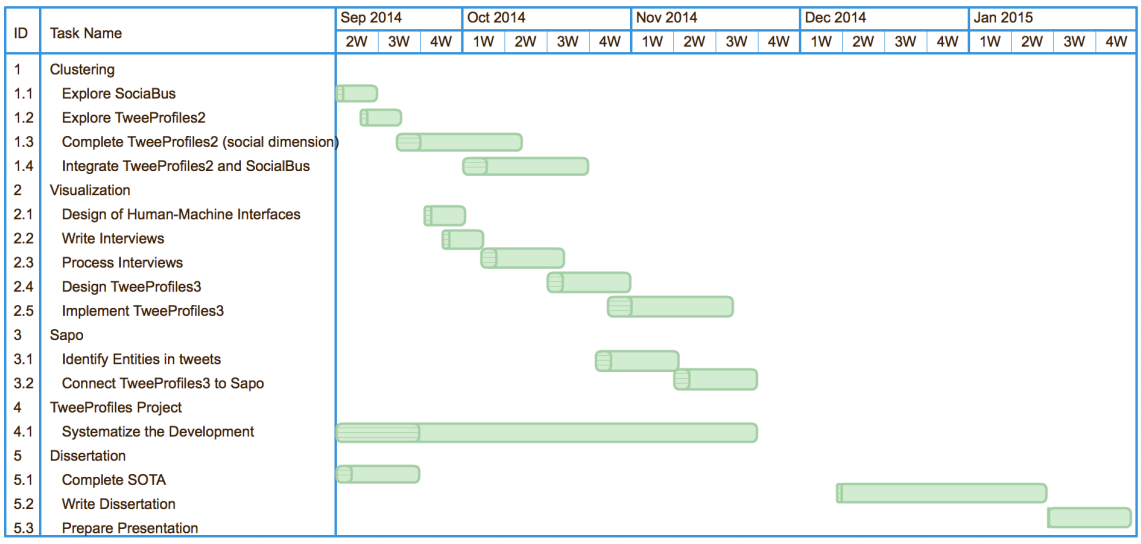


Figure B.1: Initial Gantt Diagram.

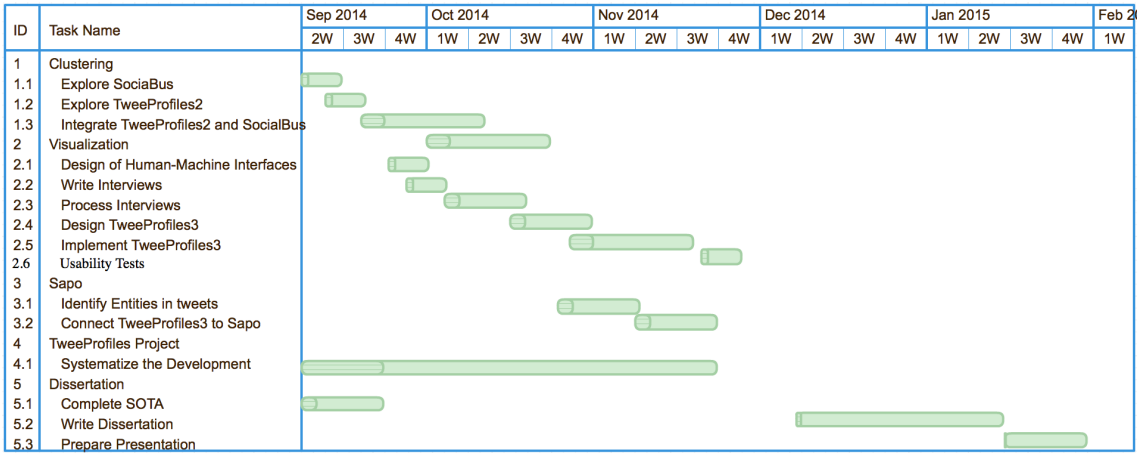


Figure B.2: Final Gantt Diagram.

References

[1] Jonathan A Silva, Elaine R Faria, Rodrigo C Barros, Eduardo R Hruschka, André CPLF de Carvalho, and João Gama. Data stream clustering: A survey. *ACM Computing Surveys (CSUR)*, 46(1):13, 2013.

[2] Joao Gama, Pedro Pereira Rodrigues, Eduardo J Spinosa, and André Carlos Ponce Leon Ferreira de Carvalho. *Knowledge discovery from data streams*. Chapman & Hall/CRC Boca Raton, 2010.

[3] Xiaotong Liu, Yifan Hu, Stephen North, and Han-Wei Shen. Compactmap: A mental map preserving visual interface for streaming text data. In *Big Data, 2013 IEEE International Conference on*, pages 48–55. IEEE, 2013.

[4] Chung-Hong Lee, Hsin-Chang Yang, Tzan-Feng Chien, and Wei-Shiang Wen. A novel approach for event detection by mining spatio-temporal information on microblogs. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 254–259. IEEE, 2011.

[5] Britney Fitzgerald. SGI Twitter Heat Map: Supercomputer Shows Where Angriest Tweeters Live. http://www.huffingtonpost.com/2012/11/19/sgi-twitter-heat-map_n_2138726.html/, 2012. [Online; accessed 01-01-2015].

[6] Maarten WIJNANTS, Adam Blazejczak, Peter QUAX, and Wim LAMOTTE. Tweetpos: A tool to study the geographic evolution of twitter topics. 2014.

[7] Alan M MacEachren, Anuj Jaiswal, Anthony C Robinson, Scott Pezanowski, Alexander Savelyev, Prasenjit Mitra, Xiao Zhang, and Justine Blanford. Senseplace2: Geotwitter analytics support for situational awareness. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 181–190. IEEE, 2011.

[8] Xiying Wang and Dan Cosley. Tweetdrops: a visualization to foster awareness and collective learning of sustainability. In *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 33–36. ACM, 2014.

[9] Andrew Crooks, Arie Croitoru, Anthony Stefanidis, and Jacek Radzikowski. # earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, 17(1):124–147, 2013.

[10] Joao Pereira. Tweeprofiles2: real-time detection of spatio-temporal patterns in twitter. Master's thesis, Faculdade de Engenharia da Universidade do Porto, 2014.

[11] Oxford Dictionaries. Oxford Dictionaries "social network". http://oxforddictionaries.com/definition/english/socialnetwork/, 2010. [Online; accessed 01-01-2015].

[12] Hohyon Ryu, Matthew Lease, and Nicholas Woodward. Finding and exploring memes in social media. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, pages 295–304. ACM, 2012.

[13] Woong-Kee Loh, Sandeep Mane, and Jaideep Srivastava. Mining temporal patterns in popularity of web items. *Information Sciences*, 181(22):5010–5028, 2011.

[14] Paolo Compieta, Sergio Di Martino, Michela Bertolotto, Filomena Ferrucci, and T Kechadi. Exploratory spatio-temporal data mining and visualization. *Journal of Visual Languages & Computing*, 18(3):255–279, 2007.

[15] Matthew A Russell. *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. " O'Reilly Media, Inc.", 2013.

[16] Twitter. About Twitter. https://about.twitter.com/company/, 2014. [Online; accessed 01-01-2015].

[17] Tiago Cunha. Tweeprofiles: deteção de padrões espácio-temporais no twitter. Master's thesis, Faculdade de Engenharia da Universidade do Porto, 2012.

[18] Tiago Cunha, Carlos Soares, and Eduarda Mendes Rodrigues. Tweeprofiles: Detection of spatio-temporal patterns on twitter. In Xudong Luo, JeffreyXu Yu, and Zhi Li, editors, *Advanced Data Mining and Applications*, volume 8933 of *Lecture Notes in Computer Science*, pages 123–136. Springer International Publishing, 2014.

[19] Ivo Mota. Olho-passarinho: uma extensao do tweeprofiles para fotografias. Master's thesis, Faculdade de Engenharia da Universidade do Porto, 2014.

[20] C Pettit, I Widjaja, P Russo, RICHARD Sinnott, R Stimson, and MARTIN Tomko. Visualisation support for exploring urban space and place. In *XXII ISPRS Congress, Technical Commission IV*, volume 25, 2012.

[21] Gennady Andrienko, Natalia Andrienko, Piotr Jankowski, Daniel Keim, M-J Kraak, Alan MacEachren, and Stefan Wrobel. Geovisual analytics for spatial decision support: Setting the research agenda. *International Journal of Geographical Information Science*, 21(8):839–857, 2007.

[22] Matko Boanjak, Eduardo Oliveira, José Martins, Eduarda Mendes Rodrigues, and Luís Sarmento. Twitterecho: a distributed focused crawler to support open research with twitter data. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 1233–1240. ACM, 2012.

[23] Jiawei Han and Micheline Kamber. *Data Mining, Southeast Asia Edition: Concepts and Techniques*. Morgan kaufmann, 2006.

[24] Hamparsum Bozdogan. *Statistical data mining and knowledge discovery*. CRC Press, 2003.

[25] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[26] Han Jiawei and Micheline Kamber. Data mining: concepts and techniques. *San Francisco, CA, itd: Morgan Kaufmann*, 5, 2001.

[27] Olfa Nasraoui. Web data mining: Exploring hyperlinks, contents, and usage data. *ACM SIGKDD Explorations Newsletter*, 10(2):23–25, 2008.

[28] Daniel Barbará. Requirements for clustering data streams. *ACM SIGKDD Explorations Newsletter*, 3(2):23–27, 2002.

[29] Feng Cao, Martin Ester, Weining Qian, and Aoying Zhou. Density-based clustering over an evolving data stream with noise. In *SDM*, volume 6, pages 326–337. SIAM, 2006.

[30] Han Jiawei and Micheline Kamber. Data mining: concepts and techniques. *San Francisco, CA, itd: Morgan Kaufmann*, 5, 2001.

[31] Alexander Boettcher and Dongman Lee. Eventradar: A real-time local event detection scheme using twitter stream. In *Green Computing and Communications (GreenCom), 2012 IEEE International Conference on*, pages 358–367. IEEE, 2012.

[32] AA Lopes, Roberto Pinho, Fernando Vieira Paulovich, and Rosane Minghim. Visual text mining using association rules. *Computers & Graphics*, 31(3):316–326, 2007.

[33] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics, 2010.

[34] John Caulfield Hannyngton. *Haversines Natural and Logarithmic Used in Computing Lunar Distances for the Nautical Almanac*. GE Eyre and W. Spottiswoode, 1876.

[35] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[36] Mark Gahegan. 11 visual exploration and explanation in geography analysis with light. *Geographic Data Mining and Knowledge Discovery*, page 291, 2009.

[37] Nacim Ihaddadene and Chabane Djeraba. Real-time crowd motion analysis. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

[38] Natalia Andrienko and Gennady Andrienko. *Exploratory analysis of spatial and temporal data*. Springer, 2006.

[39] F Bhat, M Oussalah, K Challis, and T Schnier. A software system for data mining with twitter. In *Cybernetic Intelligent Systems (CIS), 2011 IEEE 10th International Conference on*, pages 139–144. IEEE, 2011.

[40] Twitter. REST APIr. https://dev.twitter.com/docs/api/1.1/, 2014. [Online; accessed 01-01-2015].

[41] Twitter. Streaming API. https://dev.twitter.com/docs/api/streaming/, 2014. [Online; accessed 01-01-2015].

[42] James R Lewis. Ibm computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1):57–78, 1995.